Masters Program in Geospatial Technologies

SPATIAL PREDICTION OF FLOOD SUSCEPTIBLE AREAS USING MACHINE LEARNING APPROACH: A FOCUS ON WEST AFRICAN REGION

FERANMI JEREMIAH OLOWE

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School
UNIVERSITAT JAUME·I
Westfälische Wilhelms-Universität Münster

SPATIAL PREDICTION OF FLOOD SUSCEPTIBLE AREAS USING MACHINE LEARNING APPROACH: A FOCUS ON WEST AFRICAN REGION

Dissertation supervised by

Prof. Dr. Edzer Pebesma,

Institute of Geoinformatics,

Heisenbergstraße 2, 48149 Münster

Dissertation co-supervised by

Prof. Dr. Hanna Meyer,

Institute of Landscape Ecology,

Heisenbergstraße. 2, D-48149 Münster

Dissertation co-supervised by

Carlos Granell Canut, PhD

Institute of New Imaging Technologies,

Universitat Jaume I

Castellón de la Plana, Spain

DECLARATION OF ORIGINALITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all sources (published or not published) are referenced.

This work has not been previously evaluated or submitted to the Institute of Geoinformatics, WWU or elsewhere.

_____

Feranmi Jeremiah Olowe

ACKNOWLEDGEMENTS

Firstly, I express my profound gratitude to God, the giver and sustainer of life for the privilege of life and a sound mind. Additionally, I would like to appreciate my supervisors, Prof. Dr. Edzer Pebesma, Prof. Dr. Hanna Meyer and Carlos Granell (Ph.D.) for your inputs and guidance from the inception, execution, and end of this thesis. I am extremely grateful and honored to have collaborated with you, your pertinent feedback and advice were very helpful in implementing this research.

I also want to thank Professor Pedro Cabral for his unconditional advice and support whenever I needed it. I am also grateful to all the professors within the Erasmus Mundus Master Program of Science in Geospatial Technologies headed by Professors Marco Painho, Christoph Brox and Michael Gould. I also thank my classmates for their words of encouragement and assistance right from the inception of this program especially Emeka, Pablo and David Alsena, you guys are awesome.

Lastly, I appreciate my friends and family who were of great support and believed in me. Especially my parents, despite the thousands of miles away, your constant support and words of encouragement have helped me thus far. Thank you.

Spatial Prediction of Flood Susceptible Areas Using Machine Learning Approach: A Focus on West African Region

ABSTRACT

The constant change in the environment due to increasing urbanization and climate change has led to recurrent flood occurrences with a devastating impact on lives and properties. Therefore, it is essential to identify the factors that drive flood occurrences, and flood locations prone to flooding which can be achieved through the performance of Flood Susceptibility Modelling (FSM) utilizing stand-alone and hybrid machine learning models to attain accurate and sustainable results which can instigate mitigation measures and flood risk control. In this research, novel hybridizations of Index of Entropy (IOE) with Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF) was performed and equally as stand-alone models in Flood Susceptibility Modelling (FSM) and results of each model compared.

First, feature selection and multi-collinearity analysis were performed to identify the predictive ability and the inter-relationship among the factors. Subsequently, IOE was performed as bivariate and multivariate statistical analysis to assess the correlation among the flood influencing factor's classes with flooding and the overall influence (weight) of each factor on flooding. Subsequently, the weight generated was used in training the machine learning models. The performance of the proposed models was assessed using the popular Area Under Curve (AUC) and statistical metrics.

Percentagewise, results attained reveals that DT-IOE hybrid model had the highest prediction accuracy of 87.1% while the DT had the lowest prediction performance of 77.0%. Among the other models, the result attained highlight that the proposed hybrid of machine learning and statistical models had a higher performance than the stand-alone models which reflect the detailed assessment performed by the hybrid models. The final susceptibility maps derived revealed that about 21% of the study area are highly prone to flooding and it is revealed that human-induced factors do have a huge influence on flooding in the region.

KEYWORDS

Flood occurrences

Flood Susceptibility Modelling (FSM)

Human-Induced Factors

Hybrid modelling

Machine Learning (ML)

Natural-caused Factors

Remote Sensing

West African States (WASs)

ACRONYMS

ADT – Alternation Decision Trees

AHP – Analytical Hierarchy Process

ANFIS – Adaptive Neuro-fuzzy Inference System

ANN – Artificial Neural Network

ASTER – Advanced Spaceborne Thermal Emission and Reflection Radiometer

AUC – Area Under Curve

CART – Classification and Regression Trees

CHAID - Chi-squared Automatic Interaction Detection

DEM – Digital Elevation Model

DT – Decision Tree

EBF – Evident Belief Function

EM-DAT – International Disaster Database

ESA - European Space Agency

ETM – Enhanced Thematic Mapper

FR – Frequency Ratio

GI – Geospatial Technologies

GIS – Geographic Information System

IOE – Index of Entropy

LASEMA – Lagos State Emergency Management Agency

LIDAR – Light Detection and Ranging

LULC – Land Use Land Cover

MCDA – Multi-Criteria Decision Analysis

NASA – National Aeronautics and Space Administration

NDBI – Normalized Difference Building Index

NDVI – Normalized Difference Vegetation Index

NDWI – Normalized Difference Water Index

NIR – Near Infrared

OLI – Operational Land Manager

QUEST - Unbiased Efficient Statistic Tree

RBF – Radial Basis Kernel

RF – Random Forest

ROC – Receiving Operating Curve

SPI – Stream Power Index

SVM - Support Vector Machine

SWIR – Short Wave Infrared

TWI – Topographic Wetness Index

VIF – Variance Inflation Factor

WOE – Weight of Evidence

INDEX OF THE TEXT

INDEX OF TABLES

INDEX OF FIGURES

1 INTRODUCTION

1.1 Background and Motivation

Flooding is a natural phenomenon, considered as one of the major disasters overwhelming the world today because of its catastrophic and devastating after-effects[1]. Flooding is a resulting event of an excessive inundation of overland flow[2]. The hazard has been a threat to all areas of human lives in various ways such as destruction of lives and properties, wrecking of nations' economy, severe damages to transportation systems, and the alteration of biodiversity live patterns[3]. The exponential and severe economic losses caused by flooding every year globally amounts to over $20 billion with over 3000 fatalities and losses[4]. Regrettably, flooding has been labelled the costliest natural hazard because of the resulting high economic losses ranging to about 31%[5].

Similarly, the West African States (WASs) are not exempted from this cataclysmic hazard. These states are not far-fetched from the hazard due to rapid indiscriminate development and population increase[6], [7]. Based on past studies, these states are more vulnerable because of bad infrastructural planning, low level of technology, and political instability which leads to the creation of poorly resourced policies[2], [8]. However, these factors are aggravated by climatic, topographic, geomorphologic, and anthropogenic factors[9]. It has been recorded by past studies that fluvial and coastal flooding are the major flooding types that are paramount within the region which is of major concern to the urban residents and government authorities affected by this hazard[10]. According to the international disaster (EM-DAT) database, over 1,803 deaths have resulted from flooding in the last 30 years in Nigeria alone[5], [11].

Furthermore, studies have revealed that flooding occurs more in developing countries such as the WASs due to the lack of understanding and poor knowledge about flood mitigation measures and how to tackle it[12]. It makes it almost impossible to predict, mitigate, control, and manage coastal, fluvial, and flash floods[13]. Based on EM-DAT database, floods in the WASs last for 79 days on averagely and researchers have opined on the major roles of anthropogenic activities on the occurrence of floods[11], [13], [14]. It is observed that the unsystematic development within the cities, blocked drainage systems, and unsystematic waste disposal methods are a few of the anthropogenic factors that cause flooding[12]. These factors have been difficult to control because of the increasingly growing population with spaces inadequate to contain the people which leads to the construction of houses on

waterways and the filling up of water channels[15]. Consequently, a series of developments and urban expansion occurs in inappropriate locations which relate to the influence of urbanization on the reshaping of the natural environment[16].

Considering the WASs, although several policies are being established and various community-based approaches are being adopted, climate change continually fuel up the occurrence of these flood events and urbanization upsurges the threats of these events[6], [9]. As urbanization increases, so does imperviousness which increases run-off speed and exacerbates flooding[14]. The focus of these countries has always been towards the high amount of rainfall while other major factors that drive the occurrence of flooding are most times neglected. The major approach often adopted deals with the rainfall-runoff approach (hydrological modelling) which sometimes involves water percolation rate and drainage systems while other major factors are not considered because of the lack of datasets which could reveal the spatial and temporal variations of the major influencing factors of flood occurrences[13].

Consequently, there has been a high limitation and low performance of flood susceptibility modelling (FSM) in West Africa and the hydrological models utilized requires high-resolution quality datasets such as LIDAR, and heavy, complex algorithms with high computation costs in modelling flood occurrences[17]. Thus, accessibility to geospatial datasets is a paramount issue in WASs as acquiring these datasets still seems limited, and to fully obtain an accurate assessment of flooding requires the identification of flood influencing factors characteristics which are largely based on remote sensing data to develop a flood susceptibility map and identify areas prone to flooding[18]. However, there have been recent developments that have provided the platforms in acquiring these datasets such as the recently launched Sentinel satellite from the European Space Agency (ESA) and the Global ASTER DEM from NASA which gave insights into this study and has provided the opportunity for solving the problems faced in this part of the world.

Fortunately, machine learning models uplifts this burden due to its flexibility with non-linear data such as floods and has proven superiority with success in the modelling of natural hazards[19], [20]. More so, utilizing the ML approach is cost-effective and practical in data-scarce areas[3]. Furthermore, it is ascertained that creating hybrid models through the integration of statistical models with machine learning models saves time and reliable results are attained[21], [22]. Therefore, this research seeks to fill this gap by investigating not just the impact of natural factors but also the human-induced factors on flood

susceptibility and utilizing machine learning approach for modelling flood susceptibility in the region.

Considering the study's context, seven models were utilized in identifying and predicting areas prone to flooding. These models are Index of Entropy (IOE) stand-alone model, Decision Tree (DT) stand-alone model, Support Vector Machine (SVM) stand-alone model, Random Forest (RF) stand-alone model, DT-IOE, SVM-IOE, and RF-IOE hybrid models. Fifteen flood influencing factors, 139 flood locations, and 139 non-flood locations were used in training the seven models. Subsequently, flood probability indexes generated from the models were used in deriving the flood susceptibility maps. Thereafter, results attained were validated using the ROC curve and several efficient statistical indices. This study was conducted in Lagos city which is arguably the most affected city in the region affected by flood incidences (Figure 1).

Figure 1: Flooding in Lagos[23], [24]

1.2 Research Gap Identification

Flood is a global disaster that has been studied by several researchers, focusing on its mitigation and modelling. However, in recent years, through the utilization of geospatial technologies (GI) which is composed of GIS, remote sensing, and Global Positioning System (GPS), flood susceptibility mapping accuracy has increased. GI technologies have been integrated with machine learning (ML) algorithms to model areas susceptible to floods[25], [26].

To facilitate the reduction of flooding necessitates previous identification of factors influencing the occurrence of floods and areas that are highly susceptible to flood risks[22].

Accordingly, highly accurate flood susceptibility maps should be considered as an essential resource in managing flood risk. Therefore, the distinction of this research from other studies is the full consideration of factors that influences the occurrence of floods. Different studies have considered the geomorphologic, hydrological, and climatic factors which are often categorized as natural-caused factors such as slope, aspect, Topographic Wetness Index (TWI), Stream Power Index (SPI), curvature, altitude, rainfall, Land Use Land Cover (LULC), NDVI, geology and while major anthropogenic factors are mostly not considered. This is because flood influencing factors vary based on geo-environmental characteristics of the study area and factors are often selected based on existing works done in the study area[21].

However, while flooding is a cataclysmic and recurrent hazard in West Africa, no previous studies have been performed paying attention to flood susceptibility prediction. Studies have been performed emphasizing hazards description and awareness, risk, vulnerability, and feasibility studies[7], [12]. Thus, there is low knowledge and limitation on the mitigation capabilities of flood occurrences in the region. Also, these studies are conducted using secondary data sources, social and descriptive analysis as research instruments, while a few flood hydrological modelling studies performed, are conducted without a huge focus on flood influencing factors[6], [17].

Therefore, it is of paramount importance to further investigate the impact of flood influencing factors on flood occurrences to develop floodplains management approaches, lay down more refined policies, and provide more knowledge on the influencing factors. According to the United Nations, improper flood control techniques and land use, and bad urban planning practices intensify flood occurrences[27]. On a global scale, few studies have related the impact of urbanization on the occurrence of floods and an increase in runoff due to rapid urbanization, high population, enormous deforestation which drives floods occurrences.

As such, anthropogenic factors such as Normalized Difference Building Index (NDBI), population density, drainage density, distance from roads while distance from river as a natural-caused factor should be considered in flood susceptibility modelling. Past studies have related that anthropogenic factors play a huge role in the region under study which necessitates considering them in this study[28]. Also, apart from predicting the areas susceptible to flooding through the implementation of machine learning models, it is of high

necessity to identify the impact of each flood influencing factor on flood occurrence relative to the study area which is also implemented in this study.

In summary, this study attempts to assess the influence of flood influencing factors (natural-caused and human-induced) on flooding to predict future flood occurrences thereby providing key and valuable information on flood susceptibility zones and measures that can be adopted in mitigating its occurrence.

1.3 Research Aim

1.3.1 Aim

The aim of this research study is to develop and utilized machine learning-based flood susceptibility models in creating flood susceptibility maps to identify locations prone to flooding considering the human-induced and natural-caused factors while also acknowledging the impact of these factors on the area under study.

To achieve this main research aim, the following research sub-questions were addressed:

1. What is the impact of both natural-caused and human-induced influencing factors on the occurrence of floods in the study area?
2. Which machine learning technique is appropriate based on accuracy to predict areas susceptible to flooding when natural and human factors are both considered?

1.4 Methodology Overview

Within the context of the aim and research questions, the following methodology was adopted:

- Creation of geospatial database through the derivation of flood predictors (influencing factors) and flood inventory map.
- Feature selection of variables and multi-collinearity analysis to ensure each factor's predictive ability and significance.
- The implementation of bivariate and multivariate statistical model analysis adopting the IOE technique to attain each factor's influence on flood occurrence.

- Implementation of machine learning algorithms for model training using DT, SVM, and RF stand-alone models and hybrid models through the models' integration with IOE.
- Derivation of flood susceptibility maps.
- Model evaluation and performance metrics using Area Under Curve (AUC) and statistical Indices namely, Accuracy, Sensitivity, and Specificity, NPV and PPV.

1.5 Thesis Structure

- Chapter 2 (literature review) explores the past related works on flood influencing factors that drive flood occurrences and flood susceptibility modelling (FSM) approaches.
- Chapter 3 describes the study area, the preprocessing of primary datasets required to create the geospatial database for performing FSM and tools utilized for the research.
- Chapter 4 (Methodology) details the implementation procedures performed to produce the final susceptibility maps and fulfil the intents of the research.
- Chapter 5 presents the results acquired.
- Chapter 6 details the critical analysis of the results attained relating them to literature, the limitations encountered and recommendations for future steps in the research domain.
- Chapter 7 details the research's summary with its main findings presented.

2 LITERATURE REVIEW

This chapter focuses on three sections which conceptualize within the framework of this study. Section 2.1 is devoted to flood influencing factors and are further categorized into two sections based on natural and man-made driven variables. Section 2.2 details various approaches that have been utilized in the performance of flood modelling in the FSM domain. The section further details the pros and cons of each approach and how the approaches have been integrated to perform flood modelling. Finally, section 2.3 comprises various machine learning techniques that have been utilized in flood modelling with their strengths and weaknesses and weighing more on the algorithms adopted in the study.

2.1 Flood Influencing Factors

Flood influencing factors are referred to as triggers that enhances the occurrences of floods. Zhao et al. (2019) noted that the identification of influencing factors is a major procedure in flood susceptibility assessment[29]. Influencing factors are often chosen based on past related work in the study area where the most important factors have been identified as factors vary from one region to another based on the geo-environmental characteristics (topology, geology, hydrology, and anthropology) of the study area[22]. Moreover, there is no consensus on the set of influencing factors or the number of influencing factors enough for FSM[30].

Therefore, Flood influencing factors are commonly chosen based on previous studies and expert knowledge. However, it is of relative significance to acquire the geographical information related to the catchment area and its environs in flood modelling as urban catchment areas are composed of natural and artificial substances[31]. As such, each of the factors relatively important is categorized into natural-caused and human-induced and each factor is described below accordingly.

2.1.1 Natural-caused Factors

Natural influencing factors such as elevation, slope, curvature, stream power index (SPI), topographic wetness index (TWI), land use land cover (LULC), normalized difference vegetation index (NDVI), rainfall, lithology, and soil have been utilized in previous studies[18], [32]. Elevation is crucial in flood occurrence as areas with high elevation enhances an increase in runoff while flat areas are often more prone to flooding due to high water discharge[1]. Dodangeh et al. (2020) pointed out that a negative correlation exists

between flooding occurrence and elevation[33]. Slope influences surface runoff and water percolation rate into the soil as water flows from a higher elevation to a lower one[1].

Khabat et al. (2018) attested that the increase in slope causes a decrease in runoff infiltration[34]. Consequently, runoff velocity is dependent on the slope impact as steep slopes have low water percolation because of a high increase in runoff velocity[35]. Curvature is a morphometric factor that influences the occurrence of floods as divergent and convergent runoff areas are identified through curvature. It is categorized into three classes namely, concave (negative curvature), flat (zero curvature) and convex (positive curvature)[4]. It is observed that flooding occurs mostly in flat and concave areas[33]. Areas with concave and flat shapes are more susceptible to flooding as such areas retain water longer than areas with convex shape[36]. Soil characteristics differ from one region to another based on the different composition of particles which determines the level of water percolation. Its texture, type, and structure account for the runoff rate and level of water infiltration.

Furthermore, TWI describes the flow of water towards the pull of gravity within a watershed. The factor accounts for the accumulation of water in lower slope areas[16]. K. Chapi et al. (2017) defined TWI as the ratio of a specific basin area to the slope[34]. TWI identifies areas within a watershed that are prone to flooding as areas with a steeper slope have lower percolation rate unlike flat terrain[22]. TWI, therefore, indicates percolation status in a region and areas prone to flooding. SPI measures runoff's water flow erosive power[20]. Consequently, areas with highly concentrated surface runoff and high erosive power are identified[20]. It identifies the strength of flood flowing towards gravity and the amount of water accumulated in the watershed as the steeper the slope, the increase in velocity of the water flow[37]. Therefore, areas with a high tendency for flow accumulation indicates high value while low values indicate areas with low flow accumulation[38].

Also, LULC plays a vital role in flood occurrence; urban areas are more peculiar to flooding through increase runoff rate due to imperviousness of its surfaces while vegetated areas are often less flooded because of the high vegetation density. Consequently, an inverse relationship exists between vegetation density and flood occurrences[16]. Previous studies indicated that land-use patterns play a major role in flooding and should be considered in flood studies as each LULC type performs a specified role in flooding[39]. Lithology is an important factor considering spatio-temporal variation where high underlying resistant rocks or highly penetrable particles determine the drainage density rate of the area. It also

controls and determines the amount of sediment transported, speed of runoff, and percolation rate[20], [33].

Also, rainfall in most studies has been referred to as the most important influencing factor[40]. Rainfall has a remarkable influence on flood occurrence through its spatial and temporal patterns, accordingly, an increase in the amount of rainfall leads to a significant tendency of flooding[41]. Tehrany et al. (2019), noted that flooding is mainly originated by rainfall and further influenced by other factors, and the amount of rainfall drives water inundation depending on the characteristics of the basin such as its expanse, altitude and the LULC formations[22], [42]. Distance from river plays a major role in flooding and significantly determine its extent and magnitude[16]. Haoyuan Hong et al. (2018) maintained that river initiates flooding when the amount of water exceeds the amount the river network can handle[20]. Esmaeel Dodangeh et al. (2020) also emphasized that the closer it is to a river, the increase the risk of flood occurrence[33]. Kamran Chapi et al. (2017) revealed that frequent locations most affected by floods are areas close to the river[34]. Thus, it is necessary to consider distance to river as an influencing factor.

In conclusion, NDVI represents the vegetation density and indicates the vegetational characteristics of the study area, it is observed that high vegetation reduces flooding[33]. A higher NDVI increases the possibility of water percolation into the soil and reduces the possibility of flooding[19]. Consequently, a decrease in NDVI automatically increases the probability of flooding.

2.1.2 Anthropogenic Factors (Human-Induced)

Previous studies have related that population density, drainage density, normalized difference building index (NDBI), and distance from roads as anthropogenic factors play a role in flood occurrence[43], [44].

Drainage density is defined as the total steam length(m) by the total basin area (km_2) of a watershed. As a result, a high tendency of flooding in areas are identified with high drainage density[36]. Therefore, high drainage density is positively correlated with high flood peaks and volumes[34]. H. Shafizadeh-Moghadam et al. (2018), noted that drainage density describes how well-drained or poorly drained the watershed is[19]. Idowu et al. (2020) also corroborated that substandard drainage networks aggravate the occurrence of flooding and Augustine (2017) opined that poor drainage systems cause continual flooding and stream

overflow in an area[7], [12]. Also, Mahmoud et al. (2018) revealed that a dense drainage network coupled with a steep slope often leads to continual flooding[36].

Previous studies have related population density to have a significant influence on flood occurrence as an increase in population within a region causes a significant decrease in pervious spaces and an increase in urban growth[9], [15]. The Global Assessment Report on Disaster Risk Reduction (UNISDR, 2019) revealed that the high increase in population growth has caused an increase in flood risk[27]. This has been attested by previous flood modelling studies[45]. Population density is defined as the total number of people occupying a given region per unit area[46]. Hamid Darabi et al. (2019) noted that flood occurrences are highly associated with high population density and should be significantly considered in flood studies[39].

Furthermore, Distance from road influences the occurrence of floods. Gang Zhao et al. (2020) related that road enhances water inundation due to its imperviousness and forms a pathway for water flow[31]. It reduces the percolation rate which increases the runoff rate[22].

In conclusion, Normalized Difference Building Index (NDBI) indicates the building attributes within a region[33]. It a major determinant of impervious areas as the concentration of buildings increases run-off. Therefore, there exists a direct relationship between flooding and building density. Thus, should be considered as a driver of flood occurrence[38].

2.2 Flood Susceptibility Modeling Approaches

Flood Susceptibility Modelling (FSM) has been achieved through various approaches. However, each approach has its pros and cons and generates different results. Therefore, no universal consensus is laid down on model selection for FSM. Each model differ based on sensitivity to outliers, prediction accuracy, processing time and presumptions on data distribution[19]. Consequently, each of the flood modelling approaches is discussed below.

2.2.1 Hydrological Approach (Physical-based)

In the past decade, traditional hydrological and hydraulic modelling approach has been adopted by researchers for susceptibility mapping[47]. This modelling approach is categorized into three namely; one-dimensional model (MIKE 11, HECRAS), two-dimensional model (TUTFLOW, SOBEK) and three-dimensional model (Navier-Stroke)[13]. However, with this approach, fieldwork is essential and highly costly for data

gathering[47]. According to Balicia et al. (2013), highly comprehensive and detailed data is needed to achieve significant accuracy[48]. Moreover, hydrological models have been ensembled with GIS, which has proven its capability for flood modelling[47]. The model requires high computational time based on the model's dynamics and the 2D model performs more accurately than the 1D model in flood modelling but requires very long-term, high-resolution data which are burdensome to acquire and prevents short-term prediction[31]. It is revealed that most authors do not use the three-dimensional model to avoid unessential complex algorithms since some less complicated models can provide reasonable solutions[13].

2.2.2 Qualitative Approach

The qualitative approach incorporates expert knowledge and qualitative techniques to relate independent variables with flood occurrence based on numerical expressions[20]. Some of the popular qualitative techniques are the Analytical Hierarchy Process (AHP)[44], fuzzy logic[20], and Multi-Criteria Decision Analysis (MCDA)[47]. Tehrany et al. (2019), noted that qualitative models incorporate expert's opinion for its modelling considering the influencing factors and their attributes which could generate bias in the prediction modelling. The author noted that flooding is a global problem and should be predicted with an efficient and robust modelling approach[35].

However, most studies optimize this approach through the integration of the qualitative models with various decision analysis algorithms, statistical models and machine learning models[19]. Hossein et al. (2016) applied GIS ensemble method of FR and SVM to create flood susceptible mapping in Malaysia where each influencing factor is optimized by MCDA technique to generate weights that serve as inputs for SVM model[47]. Rahmati et al. (2016) utilized the MCDA technique for flood modelling and generated flood susceptibility maps[49]. However as opposed by Tehrany et al. (2019), the method is unsuitable for flood susceptibility studies as expert knowledge is integrated into the model, it was further attested by Romulus et al. (2020), who utilized fuzzy AHP model which achieved low performance based on expert's judgement involved in the modelling[4].

Rahmati et al (2016) however noted that AHP is simple, budget-friendly, less time-consuming and easier to develop for flood susceptibility studies and more suited for regional studies[47]. Samantha et al. (2018) also adopted MCDA which was conducted and compared with the FR model. The author related that FR model had a better performance than MCDA[25]. Hong et al. (2018) adopted the fuzzy logic technique for FSM and revealed

that the technique chosen does not quantify the variables' importance as the variables were chosen based on expert judgement on the study area. He further argued that incorporating field experience and expert judgements generate more accurate results[20].

2.2.3 Statistical Approach

The statistical approach has been utilized by researchers in recent times, some of these methods are frequency ratio (FR)[21], logistic regression(LR)[50], weight of evidence (WOE)[51], Index of Entropy (IOE)[52] and Evident Belief Function[22]. This approach works on a presupposition that flood influencing factors are associated with, and drives the occurrence of flood events[1]. Tehrany et al. (2019) opined that most statistical methods rely on linear presumption while flood is a multidimensional phenomenon, the author related that ensemble statistical methods augment this flaw such as the adaptive neuro-fuzzy inference system (ANFIS) however, this model requires various parameters to perform FSM accurately[35].

This was counteracted by Pradhan et al. (2015) who LR and noted that LR utilizes both continuous and discrete variables in FSM thereby describing the flexibility of the model[30]. Statistical modelling techniques are categorized into bivariate and multivariate statistical models. Bivariate statistical models such as FR and WOE are probabilistic models which measure the occurrences of flood based on each class of the influencing factors[45]. Therefore, the bivariate probability is calculated by correlating each class of the influencing factors with flood occurrence and the higher the bivariate probability the stronger the impact of that factor on flood occurrence[53]. However, it is noted that the bivariate approach is based on generalization as interaction among factors are not considered and weights not assigned to each factor[16]. This presumes that flood occurrence is based on the same set of factors with equal weights across the study area[21], [26].

On the other hand, multivariate statistical model such as LR correlates each influencing factor directly with flood occurrence and performs correlation among the influencing variables[50]. However, the weak point of statistical models is its inefficiency in handling complex and multidimensional phenomena because of its linear structure[54]. Moreover, It is noted that the combination of FR and LR increases the efficiency of the model and cover up the weakness of both models[30]. Furthermore, hybrid models formed from the integration of statistical and machine learning models have been ascertained by past studies to yield more accurate results[4]. According to Hong et al. (2018), hybrid model's flexibility allows extensive assessment of influence on each flood-related independent variables in

each class[20]. Wei Chen et al. (2020) related that increase in accuracy of flood studies requires the precise identification and greater prediction capabilities in forecasting future flood occurrences which is achievable through the hybridization of statistical and machine learning models[40].

2.3 Machine Learning Modelling Approach

Machine learning (ML) is an efficient approach to the prediction of natural hazards[55]. The approach helps to detect and uncover insights, relationships among variables which makes it suitable for predictive modelling. Various methods have been applied to flood susceptibility modelling (FSM). Models such as random forest (RF)[38], decision trees(DT)[21], support vector machine(SVM)[22], artificial neural network (ANN)[56], multivariate adaptive regression splines (MARS)[33], neuro-fuzzy inference system (ANFIS)[54], have been utilized for FSM studies. These models have been compared with their performances noted. However, to date, there is no superiority among the models as each has its pros and cons[26]. In recent times, ML approaches are being integrated with geospatial technologies in flood mapping studies for handling more complex phenomena and computing large amounts of data accurately[54]. ML techniques have the advantage of predicting and modelling complex structures in a proficient manner[16].

Furthermore, in recent times hybridization of machine learning models are being adopted. The aim of utilizing the hybrid models is to increase the predictive capability and precise identification of areas susceptible to flooding[4]. Thereby, the influence of the conditioning factors on flooding can be detailly assessed. According to Amir Morsavi et al. (2018), the hybridization of machine learning models enhances performance and increases accuracy[55]. Romulus et al. (2019) noted that no single method is appropriate for flood modelling and revealed that the utilization of hybridized machine learning models eradicates the weaknesses of ML models and generates more accurate results[52].

ML approach learns the relationship between the flood influencing factors and flooding occurrences without subjecting to an expert opinion which often leads to bias[38]. ML models are advantageous in assessing any kind of data type (categorical, nominal, and continuous) which makes the algorithm more flexible[43].

However, as revealed by researchers in various instances, ANN and ANFIS have been adopted in flood studies, and both techniques are robust in the presence of outliers and efficient in handling errors in the input dataset. However, it is noted that these algorithms

are difficult to understand and implement even with high versatility on incomplete datasets[35]. Sayed Tameh et al. (2018) supported this fact as tuning of function parameters complicates the utilization of the algorithm and noted that it is essential to optimize the parameters with other algorithms to increase its flexibility[16]. A brief theoretical background and overview of a few notable ML algorithms are detailed below.

2.3.1 Decision Tree (DT)

DT is an efficient ML technique that has performed effectively in flood modelling[3]. Its procedural approach is easy to create and interpret. Although decision trees consume a lot of time in classifying and computing, the algorithm can deal with uncertainties to a significant extent in a dataset[35]. The algorithm is flexible in handling data with various scales, assumes no statistical distribution and has high efficiency in creating rules for predicting complex relationships[16]. DT algorithm classifies the influencing factors in a hierarchical manner and equivalently in accordance with the susceptibility levels, and create decision rules based on an established tree structure built on the significant levels of the set of independent parameters utilized[35].

Thus, the set of parameters are analyzed to generate an outcome. Decision trees have been integrated with different algorithms such as the naïve bayes tree (NBTree)[3] and alternating decision tree (ADT)[18] and other processing techniques such as the Classification and Regression Trees (CART)[51], Chi-squared Automatic Interaction Detection (CHAID)[57], Unbiased Efficient Statistic Tree (QUEST)[39]. Each of these DT classifiers has been used in modelling studies and each has its pros and cons. The NBTree is often combined with the J48 algorithm to increase the predictive capability of the algorithm[34].

2.3.2 Random forest (RF)

RF is a classification and regression modelling approach. RF is based on a fusion of random subspace method and bagging ensemble learning[38]. RF algorithm combines decision trees in predicting an outcome by permuting each variable randomly and the prediction results acquired is compared with each variable to obtain its significance. In this context, a training dataset D = ((A1, B1), …, (A_n ,B_n)) which contains the n vectors. A ∈ X and B ∈ Y where X represents numerical observations and Y represents the class labels (flood and non-flood)[3], [58]. RF works based on two processes namely bagging and random selection. This is performed to prevent overfitting within the dataset and to increase the predictive ability of the model. The process is popularly referred to as the out-of-bag procedure where

samples within the dataset are randomly drawn and replaced till a most minimal error is achieved[14]. RF is versatile in handling inconsistent and missing data and performs well with multi-collinearity and the performance of the model is based on the number of decision trees ($Ntree$) and variables attributes in the subset ($Mtry$)[34]. An increase in the value of $Ntree$ increases the time consumed in modelling while the modelling becomes more prone to errors when the $Ntree$ is small[59].

2.3.3 Support Vector Machine (SVM)

SVM is a high predictive model with high versatility, till recent, SVM has not been explored in FSM[55]. It relies on a statistical learning approach that generates output values from a number of input values[22]. It is a supervised machine learning technique that converts non-linear structures into linear by generating a hyperplane to simplify and distinguish classes in the data while categorizing the data into the training and testing dataset[35]. SVM, based on predictive accuracy, is suitable for FSM as it handles data independently of the measurement scales and works efficiently with any data format[22]. Hong et al. supported this fact and revealed the effectiveness of SVM in classifying linear and non-linear data[20].

Researchers, however, noted that the efficiency of SVM performance highly depends on the kernel adopted and the influencing factors adopted[30]. The SVM kernels mostly used are linear kernel (LN), polynomial kernel (PL), radial basis function (RBF), and sigmoid kernel (SIG)[35]. Radial basis function kernel (RBF) has often been implemented in previous studies because of its efficiency and high accuracy. Hong et al. (2018) noted that SVM is disadvantageous because of its inability to measure the significance of attributes chosen[20]. However, this drawback was resolved by Tehrany et al. (2019) who utilized kappa index with the SVM model to detect the significance of the attributes[22]. A kernel function is used in distinguishing and transforming the data[21]. Afterwards, the main input of the training dataset is mapped into a high dimensional space where the split hyperplane is created in the original space of n coordinates to distinguish between points of two different classes (flood, non-flood)[22].

According to Tehrany et al. (2019), several hazard modelling studies have revealed that DT is moderately robust than SVM in modelling natural hazards[35]. However, other studies reveal that both ML algorithms are both efficient and robust in hazard modelling and offer similar results.

3. DATA AND CASE STUDY

The following chapter describes the area of study, datasets and tools utilized in the study. Furthermore, it focuses on the geographical derivations and the creation of the geospatial database that was used in performing the flood modelling. Section 3.1 defines the study area while section 3.2 details the preprocessing and the creation of the flood inventory dataset, section 3.3 details the flood influencing variables derivation and analyses majorly from satellite imageries using ArcGIS Pro's spatial analysis and spatial statistical tools.

3.1 Case Study

The study focuses on the entire region of Lagos state, a metropolitan coastal low-lying city located in the South-western region of Nigeria, West Africa. It is regarded as the economic hub of Africa. The city's geographical coordinate ranges between 3.1º to 3.4º E longitude and 6.5º to 6.8ºN latitude[12]. The city experiences an equatorial (humid and hot) climate with a double-maxima rainfall all through the year. The region's climate has two distinct periods namely the rainy season (April – October) and the dry season (November – March). The city is composed of mangrove swamp and forests where the mangrove swamps dominate the south while the forests are majorly found in the northern areas of the region[7].

The city is composed of highly dense road networks with inland waterways and her southern boundaries are defined by about 180 km of Atlantic coastline and a border along the western perimeter with the Republic of Benin. The city has a total landmass of about 2,345km2 which represents about 0.4% of Nigeria's total land area[10]. It is a highly populated region, and its population density continues to increase due to its commerciality. Based on its rapid urbanization, there has been a series of expansion towards the creeks and lagoon within the city[28]. Recurrently, flood occurs within the city with a destructive impact on lives and properties which forces the evacuation of people from their residences. Thus, regarded as one of the West-Africa cities highly prone to flooding.

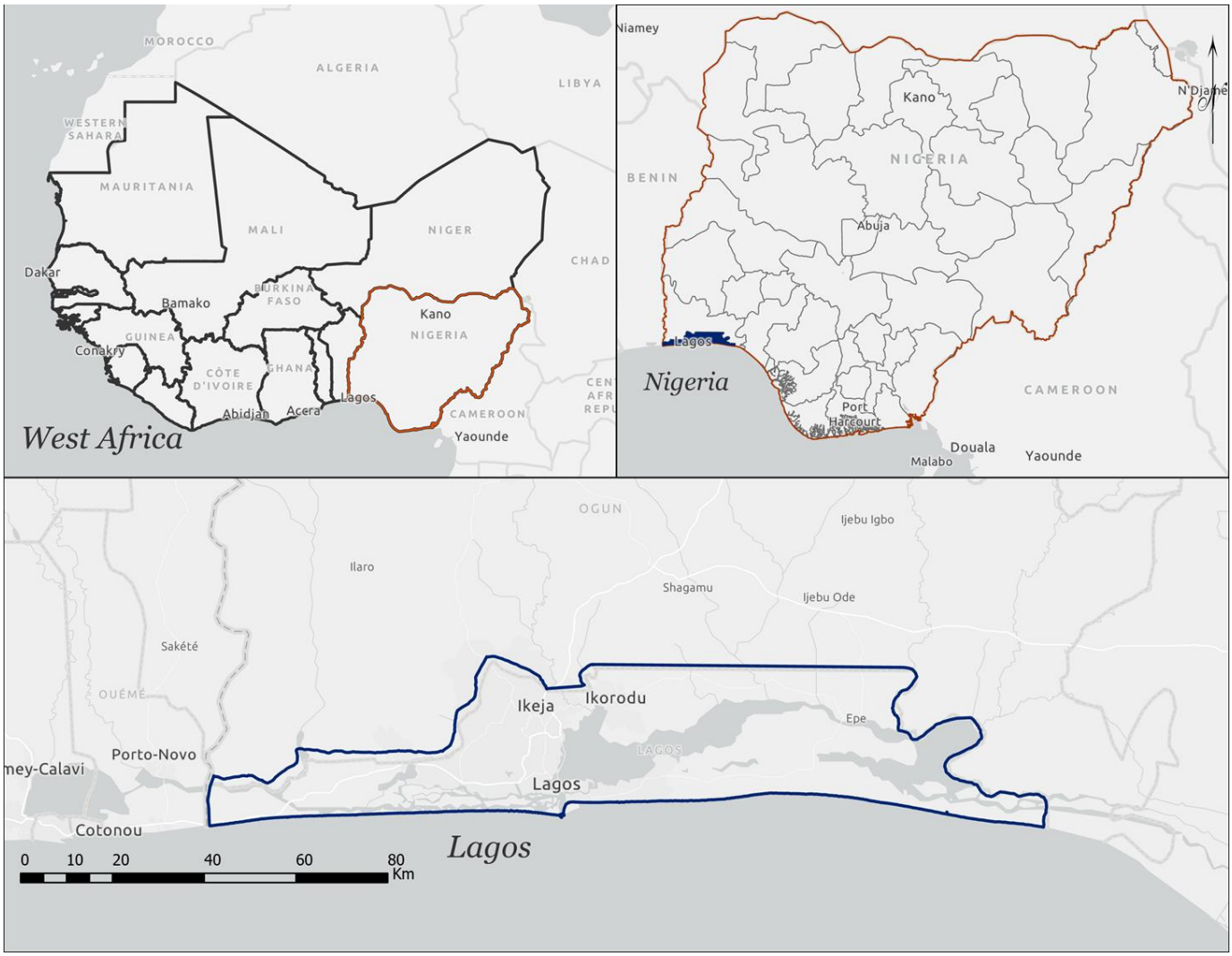


Figure 2: Case Study

3.2 Datasets and Preprocessing

3.2.1 Derivation of Flood Inventory Dataset

Flood susceptibility mapping demands two sets of data. The first dataset constitutes the past flood locations (flood inventory) which indicate the past flooded regions while the second dataset constitutes the flood triggering parameters otherwise referred to as the flood influencing factors[32]. Flood susceptibility is based on the impact of the flood influencing factors on the occurrence of floods which entails assessing the significance of each contributing parameter on flood occurrence. In this study, the flood inventory dataset was provided by the Lagos State Emergency Management Agency (LASEMA) which comprises past flood location events within the study area from 2010 to 2020 and was augmented with

satellite imageries, flood historical records and data derived from the international disaster (EM-DAT) database. A total of 139 flood events were located within the study area and an equal number of 139 non-flood points was created across the study area using the 'create random point' tool. To verify the correctness of the flood locations and non-flood locations, the Normalized Difference Water Index (NDWI) was calculated from Landsat 7 ETM+, Landsat 8 OLI, and Sentinel 2 satellite imageries using the near-infrared (NIR) and short-wave infrared (SWIR) bands to identify the past flood locations from 2010 to 2020. It is calculated as follows:

$$\mathbf{NDWI} = \frac{NIR - SWIR}{NIR + SWIR} \quad (1)$$

The flood locations were represented as points as polygon formats yields exaggerated results and become complicated for the algorithms utilized[50]. Therefore, the points were pinpointed on the centroid of the flooded areas. This has been further proven by hazards modelling studies which utilized the point format for flood inventory and generated accurate results[20]. The flood inventory map (Figure 3) was further divided into training and testing dataset as required for the training and validation stages, respectively. There is no consensus on how the inventory data is classified as it is highly dependent on the availability and quality of data. Space robustness and time robustness are two standards for classifying flood inventory. Time robustness involves dividing flood inventory data into two periods of past occurrence and future occurrence which represents training and testing datasets, respectively[22].

However, acquiring temporal data is burdensome as each flood occurrence is associated with the precipitation that triggered it and goes for other spatio-temporal influencing parameters. In space robustness, flood inventory data is randomly divided into two datasets namely training and testing datasets [35]. In this study, both standards were integrated based on the flood inventory data available. Therefore, the flood inventory data was divided into 70% for training and 30% for testing based on the 2010 to 2020 flood data. Flooding is a binary classification modelling; therefore, it was required to create equal 139 non-flood location points to ensure consistency and accuracy. Consequently, the 139 flood and 139 non-flood locations were divided using random selection technique into training and testing datasets, respectively. Values 0 and 1 were assigned to the non-flood and flood points respectively, where 0 represents flood non-existence and 1 represents flood existence.

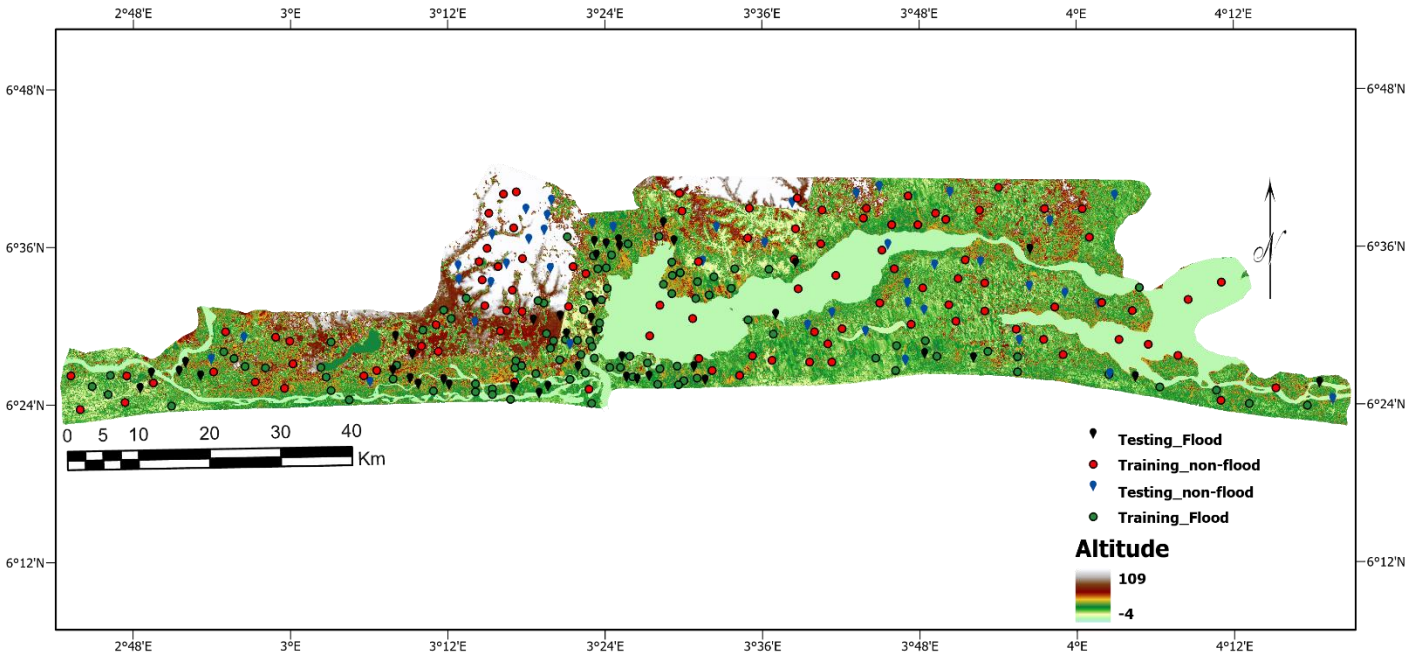


Figure 3: Flood Inventory Map

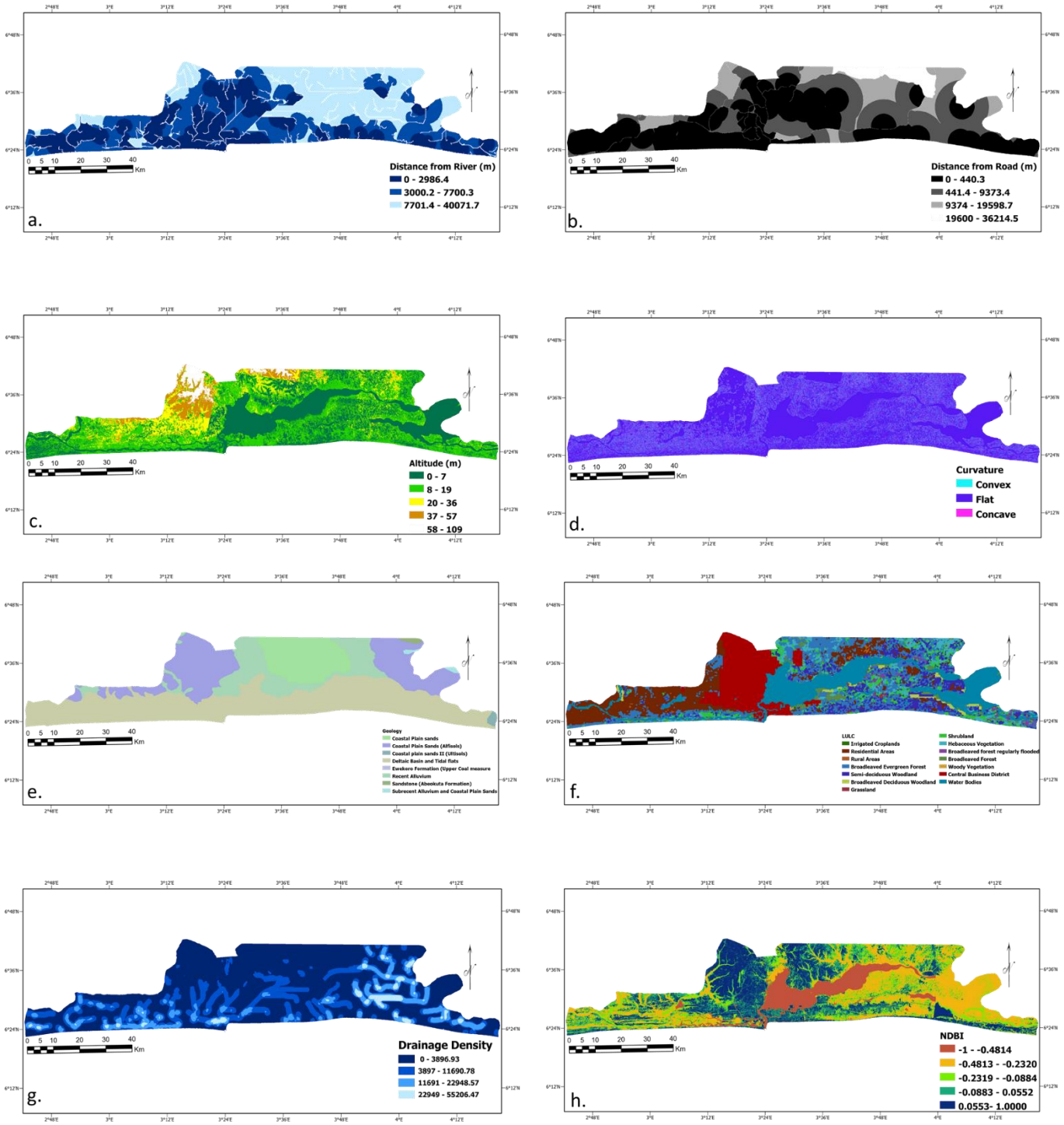
3.2.2 Derivation of Flood Influencing Factors Datasets

Satellite imagery has successfully proven to play a major role in hazard assessment based on its combination of spatial, spectral and temporal resolution which identifies past and present occurrences of hazards[4]. Sentinel 2 imagery was utilized in this study to derive the flood predictors namely LULC, NDBI, NDVI. The Sentinel-2 imagery provided a spatial resolution 10m-60m), multi-spectral features (13 bands) and temporal resolution (five days with two satellites at the equator)[60]. The LULC was obtained from the classification of the Sentinel 2 imagery using the maximum likelihood algorithm which achieved an overall accuracy of 89%. The Sentinel-2 imagery was acquired on October 15, 2020, and a spatial resolution of 10m and four spectral bands; red (B4), green (B3), blue (B2) and NIR (B8) were used in deriving the LULC map. Google Earth imagery was used as the training data and a total of 14 LULC classes were distinguished within the study area.

On the other hand, the NDVI and the NDBI were also calculated using the SWIR (B11) bands, NIR bands and red bands. The formula used for calculating both indices are as follows:

$$\mathbf{NDBI} = \frac{SWIR - NIR}{SWIR + NIR} \qquad (2)$$

$$\mathbf{NDVI}=\frac{NIR-Red}{NIR+Red} \quad (3)$$



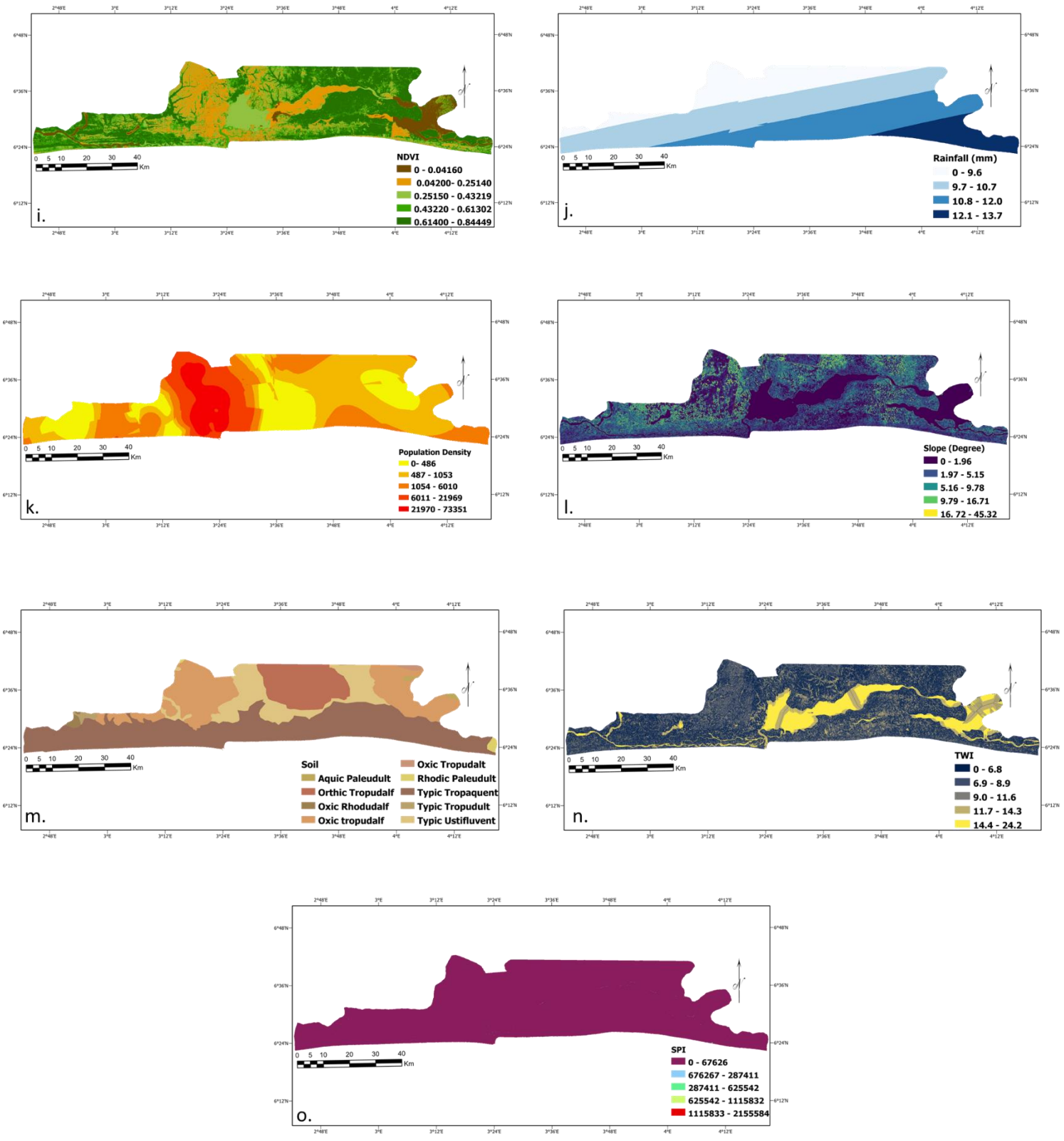


Figure 4: Flood influencing factors; (a) Distance from River, (b) Distance from Road, (c) Altitude, (d) Curvature, (e) Geology, (f) LULC, (g) NDBI, (h) Drainage Density, (i) NDVI, (j) Rainfall, (k) Population density, (l) Slope (m) Soil, (n) TWI, (o) SPI.

Furthermore, a digital elevation model (DEM) for the watershed was obtained from the advanced spaceborne thermal emission and reflection radiometer (ASTER Global DEM) 30m database. The following flood predictors such as elevation, curvature, slope, TWI, SPI, distance from river (river network), and drainage density were derived from the DEM (Table 1).

Influencing Factor	Source of Data	Data Type	Scale and Resolution	Data Source
Rainfall	UEA climatic research unit	GRID	30m	https://sites.uea.ac.uk/cru/data
Altitude	Derived from DEM	GRID	30m	https://earthdata.nasa.gov/learn/articles/new-aster-gdem
Curvature	Derived from DEM	GRID	30m	"
Slope	Derived from DEM	GRID	30m	"
TWI	Derived from DEM	GRID	30m	"
SPI	Derived from DEM	GRID	30m	"
Drainage Density	Derived from DEM	GRID	30m	"
Distance from Road	Geofabrik Website (Open Street Map)	Line Coverage	30m	https://www.geofabrik.de/
LULC	Classifying Sentinel-2 Imagery.	GRID	30m	https://scihub.copernicus.eu/
Soil	Digital Soil map (DSM) database	Vector	1:250,000	https://data.mendeley.com/datasets/zmrt6k83wk/draft?a=d9a35c1e-c19b-4ddd-b34e-69674a8ceb18
Geology	Digital Soil map (DSM) database	Vector	1:250,000	https://data.mendeley.com/datasets/zmrt6k83wk/draft?a=d9a35c1e-c19b-4ddd-b34e-69674a8ceb18
Population	City population Website	GRID	30m	http://www.citypopulation.de/
Distance from River	Derived from DEM	GRID	30m	https://earthdata.nasa.gov/learn/articles/new-aster-gdem
NDVI	Derived from Sentinel-2 Imagery	GRID	30m	https://scihub.copernicus.eu/
NDBI	Derived from Sentinel-2 Imagery	GRID	30m	https://scihub.copernicus.eu/

Table 1: Spatial Datasets and Data sources

Curvature was classified into three classes namely concave (positive value (+)), flat (value 0) and convex ((negative value (-)). The spatio-temporal factors such as the SPI and TWI was calculated from as DEM as follows:

$$\mathbf{TWI} = \mathbf{In}\ (As/tan\ \beta) \quad (4)$$

$$\mathbf{SPI} = Astan\ \beta \quad (5)$$

Where As equals the area of the catchment($m2$), and β (radians) equals slope gradient[22]. Furthermore, the drainage density was derived using the following equation stated below:

$$\mathbf{DD} = \frac{1}{S}\sum_i L\frac{s}{i} \quad (6)$$

Where S equals the study area and L_i^S equals the length of the river within the study area[36]. The population data was obtained from the city population website[61] which was processed using an areal interpolation tool to derive the population raster map. The mean annual rainfall data from 2010-2020 was obtained from a high resolution spatial gridded dataset provided by the climatic research unit, university of East Anglia[62]. The road data was obtained from the open street map through the Geofabrik website[63], while the river network was derived from the DEM using the flow accumulation and flow direction tool and distance to both attributes were derived using the Euclidean distance tool.

Soil and Lithology at a scale of 1:250,000 were extracted from the free open adaptable digital soil map database of Nigeria created by Ugonna et al(2020)[64]. The Soil was categorized into eight classes identified as Oxic Rhodudalf, Rhodic Paleudult, Oxic Tropudalf, Typic Tropudult, Typic Tropudult, Oxic Tropudalf, Aquic Paleudult, Orthic Tropaudalf and Typic Tropaquent while lithology is categorized into nine classes identified as Coastal plain sand (Alfisols), Coastal plain sands, Recent Alluvium, Sandstone (Abeokuta Formation), Transitional materials of subrecent alluvium, Coastal plain sands II (Ultisols), Ewekoro Formation(Upper coal measure), and Deltaic Basin and tidal flats respectively. All the other predictors were reclassified to the desired classes using the quantile classification method to create the flood database. The flood influencing factors were all re-processed to a 30 by 30m pixel size that corresponds to the DEM's spatial resolution and re-projected to the UTM zone Minna 31N to create the geospatial database.

4. METHODOLOGY

The methodological approach for this research study is divided into six steps and procedurally detailed as follows: (A) Data derivation, pre-processing and preliminary analysis as laid out in chapter 3. (B) Feature Engineering (Feature Selection and Multi-collinearity analysis). (C) Data Cleaning and Normalization. (D) Index of Entropy bivariate modelling (E) Machine learning models training and the generation of flood susceptibility maps (F) Models validation using Area under Curve and other statistical indices. A detailed description of each step is presented in the following subsections (Figure 5).

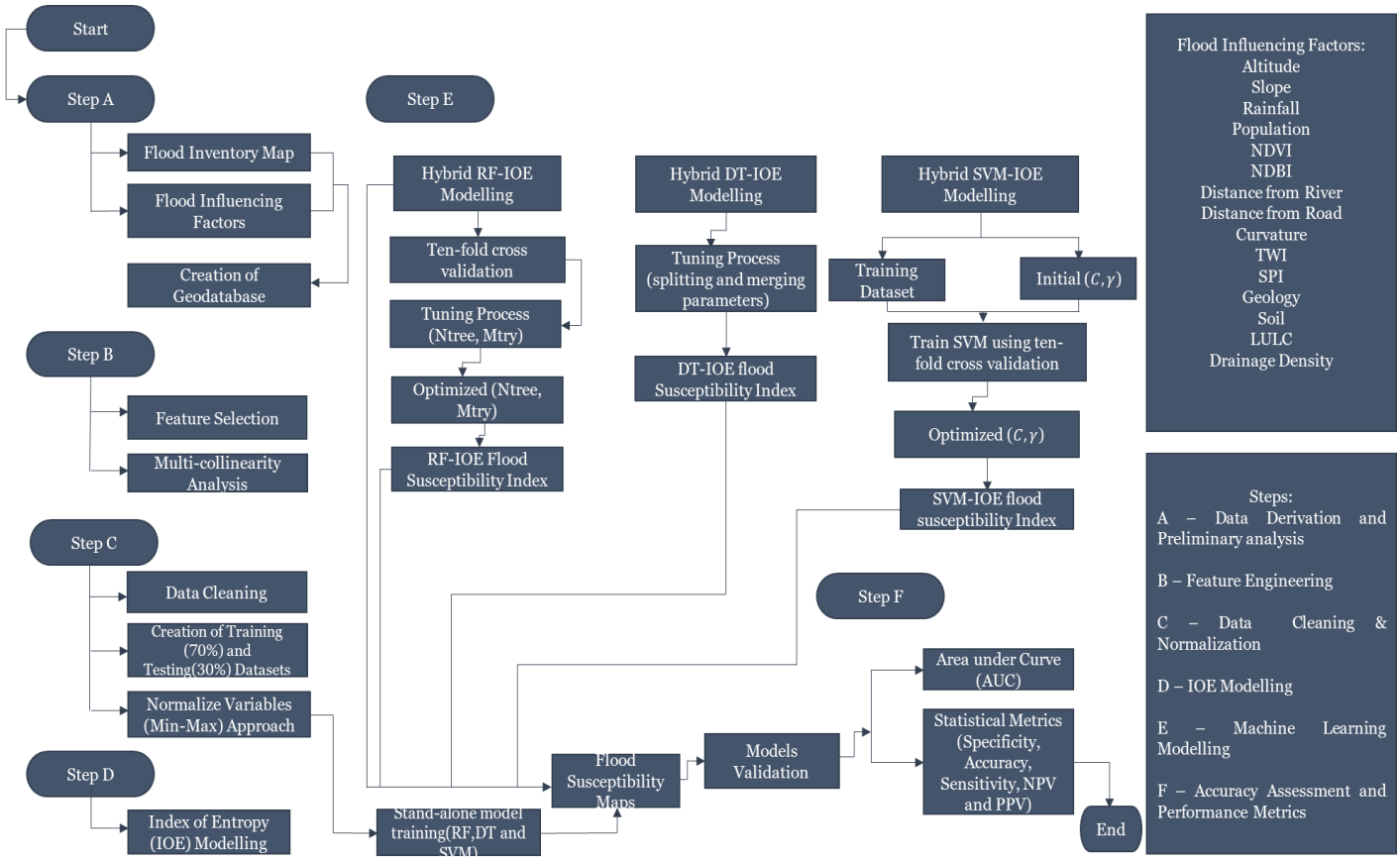


Figure 5: Methodology Flowchart

4.1 Feature Selection and Multi-collinearity Analysis

This is a process to identify the predictive capabilities of the flood predictors as factors selected fully depends on the geo-environmental characteristics of the study area and factors that have low predictive power which could generate outliers and decrease the model's predicting ability can be identified and removed[4]. To ensure accuracy as no previous studies have been done in the region, linear support vector machine was utilized in checking

each factor's predictive ability and significance (Step B). Also, to avoid redundancy among the factors, multicollinearity analysis was performed.

4.1.1 Linear Support Vector Machine (LSVM)

This is an efficient means of evaluating a predictor's capability. It has been proven to enhance the classification accuracy in modelling[4], therefore considered for this study. Considering the training dataset and the fifteen influencing factors, the LSVM equation is calculated as follows:

$$\mathbf{f(x) = sign\ (}w^T\ \mathbf{a + b)} \quad (7)$$

Where f(x) represents the function upon which the linear support vector machine is derived, w^T represents the inverse matrix of weight associated with each flood influencing factor, a = (a1, a2,….a15) represents the input vector that contains the flood influencing factors, and b represents the offset of the hyperplane's origin[22]. The factors are selected based on the average merit for each factor which ranges from 0 – 1. Consequently, a factor equal to zero is excluded from the analysis. The LSVM algorithm was implemented in the WEKA environment using the attribute evaluation tool.

4.1.2 Multi-collinearity Analysis

This is to check if two or more influencing factors are highly correlated which could cause redundancy and reduce the accuracy of the models utilized for the study[20]. All factors introduced are very important in terms of anthropogenic, climatic, hydrological, and geomorphological unique characteristics. Consequently, Tolerance (TOL) and Variance Inflation Factor (VIF) approaches was utilized where VIF > 10 and TOL < 0.1 standard indicates multicollinearity in the influencing factors as laid out in existing literature[35]. This approach has proven its efficiency in hazard modelling with a success record. The VIF measures the correlation among variables by inflating the variance of each variable within the regression's coefficient while TOL is the inverse of VIF[20].

4.2 Data Cleaning and Normalization

After the creation of the geospatial database, data cleaning and normalization was required before performing the flood modelling (Step C). Therefore, seven missing values found in the database were replaced by the mean of each variable by computing differential statistics for each of the influencing variable. Also, to avoid varying scales among the variables which could alter the results of the machine learning algorithm, the min-max normalization

approach was utilized with values ranging between 0 and 1. The normalization equation is represented as follows:

$$\boldsymbol{X'} = \frac{(\boldsymbol{X}-\boldsymbol{X_{min}})}{(\boldsymbol{X_{max}}-\boldsymbol{X_{min}}} \quad (15)$$

Where X represents the original value, X' equals the normalized value, and X_{min}, X_{max} represents the minimum and maximum values for each influencing variable[60].

4.3 Index of Entropy Modelling

The index of Entropy is a bivariate statistical approach popularly utilized in the modelling of natural hazards[53]. Previous studies have utilized this method and generated highly accurate results in susceptibility modelling[51]. Consequently, the approach was adopted in this study (Step D). Index of entropy measures variabilities and instabilities within a database. Considering this study, the extent to which flood influencing factors triggers flood occurrence is represented by the entropy of a flood event thereby generating the weight of each influencing factors. Therefore, the weight of each influencing factor in the flood probability index will be ascertained. To perform IOE, each influencing factor was classified using the quantile classification technique to ensure even distribution of pixels across each class and a reliable assessment of each class's impact on flood occurrence is identified. Furthermore, the IOE model generates two main outputs namely weights associated with each factor and each factor's classes. Each factor's weight (W_j) is calculated as follows:

$$(\boldsymbol{P_{ij}})=\frac{\boldsymbol{FR_{ij}}}{\sum_{j=1}^{S_j}\boldsymbol{FR_{ij}}} \quad (8)$$

Where FR_{ij} represents the frequency ratio coefficient for each class of each influencing factors; S_j represents the number of classes, and (P_{ij}) represents the probability density[4].

$$\boldsymbol{H_j}=\sum_{\boldsymbol{i=1}}^{\boldsymbol{S_j}}(\boldsymbol{P_{ij}})\ \boldsymbol{log_2}\big(\boldsymbol{P_{ij}}\big), \boldsymbol{j = 1, 2, \dots, n} \quad (9)$$

$$\boldsymbol{H_{jmax}} = \boldsymbol{log_2}(\boldsymbol{S_j}) \quad (10)$$

$$\boldsymbol{I_j} = \frac{\boldsymbol{H_{jmax}}-\boldsymbol{H_j}}{\boldsymbol{H_{jmax}}}, \mathbf{I = (0,1)}, \boldsymbol{j = 1, \dots, n} \quad (11)$$

$$\boldsymbol{P_j} = \frac{\boldsymbol{1}}{\boldsymbol{S_j}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{S_j}}\boldsymbol{P_{ij}} \quad (12)$$

$$\boldsymbol{W_j} = \boldsymbol{I_j} * \boldsymbol{P_j} \quad (13)$$

Where H_j *and* H_{jmax} equals the values of entropy; I_j represents the information coefficient; P_j represents the empirical probability; W_j equals the weight values associated with each flood influence factor.

Considering the stand-alone model, flood probability index final values is calculated, and the equation is as follows:

$$\boldsymbol{FSI_{IOE} = \sum_{i=1}^{n} \frac{Z}{m_j} * C * W_j} \quad (14)$$

Where i equals the total number of conditioning factors; Z equals the number of classes of the factor having the highest number of classes; m_j represents the number of classes of each factor; C equals the calculated rate of each class; and W_j represents the final weight of each factor. The advantage of IOE is it can be used for both BSA and MSA[52] which was implemented in this study. The final weights derived from each factor were used as inputs in modelling the hybrid machine learning models.

4.4 Machine Learning (ML) Algorithms

In this subsection, the three ML model's implementation in this study is briefly explained. The ML models are adopted to identify the correlation existing between the influencing factors and flood occurrences and to forecast flood susceptibility in the study area (Step E). Therefore, SVM, RF, and DT were utilized, and the implementation of each ML algorithm is described below:

4.4.1 Support Vector Machine

The training dataset contains instance-label pairs (x_i, y_i), with $y_i \in R^n$, $y_i \in \{1, -1\}$, and i =1, …, m. x signifies the vector within the input space which incorporates all the influencing factors (elevation, slope, curvature, rainfall, geology, soil, LULC, NDBI, NDVI, TWI, SPI, population density, drainage density, river, and roads). SVM sets up an optimal hyperplane that distinguishes and separates flood and non-flood pixels into {1, 0} in the training set. Separating a hyperplane using linear separable data is defined as:

$$\mathbf{y_i(w.x_i + b) \geq 1 - \varepsilon_i} \quad (16)$$

Where w represents the coefficient vector through which the hyperplane's orientation is defined in the feature space, b represents the hyperplane's offset from its origin and $\varepsilon_i$represents the positive slack variable. Lagrangian multipliers are solved to find an optimal hyperplane. It is calculated thus:

$$\sum_{i=1}^{n} \propto_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \propto_i \propto_j \; y_i y_j (x_i x_j), \quad (17)$$

$$\sum_{i=1}^{n} \propto_i y_j = 0, \; 0 \leq \propto_i \leq C, \quad (18)$$

Where $\propto_i$ are Lagrange multipliers, C is the penalty and ε_i represents the slack variables that allow the penalized constraint violation. The step-by-step layout of SVM modelling is well described in Tehrany et al (2014)[65]. The decision function of SVM classification is defined as:

$$g(x) = sign\{\sum_{i=1}^{n} y_i \propto_j K(x_i . x_j) + b\} \quad (19)$$

Where K (x_i, x_j) represents the kernel function. The kernel function is mathematically expressed as:

$$K(x_i, x_j) = exp(-\gamma / / x_i - x_j ||^2) \quad (20)$$

The penalty (C) and the kernel width (γ) were optimized using the cross-validation approach to actualize accurate results as opposed to the trial-and-error technique to prevent overfitting. To perform cross-validation requires dividing the training dataset into n folds where one-fold is reserved for testing and the remaining folds (n-1) fold is used for training. The average accuracies of the validation are established and used in generating the final flood susceptibility model. Therefore, ten-fold cross-validation was used in this study by dividing the training dataset into 10 random groups till the best values of these parameters were actualized. The optimized parameters (C, γ) were used in generating the final SVM and SVM_{IOE} probability maps. RBF kernel was utilized in this study as its efficiency and advanced interpolation and extrapolation potentialities have been proven in hazard modelling studies and various literature sources[22].

4.4.2 Random Forest Model

The random forest modelling depends on the user-defined parameters namely as the m_{try} and n_{tree}. The m_{try} represents the number of variables randomly selected at the split of each node and n_{tree} represents the number of trees contained in the model[20], [34]. Therefore, the m_{try} values are within the range of [1, 15] while the n_{tree} values are within [500, 1000, 1500, 2000, 2500] range. Ten- fold cross validation method was also utilized in tuning these parameters to ensure their values are within the number of flood variables range. The tuned parameters actualized were used in creating the final RF and RF_{IOE}probability indexes.

4.4.3 Decision Tree Model

The chi-squared Automatic Interaction Detection (CHAID) algorithm was chosen for the flood susceptibility modelling as each factor used in building a branch shows a strong correlation with the dependent variable and each new branch is created based on the relative importance relationship between the influencing factors and flood inventory while factors signifying the same influence are consolidated to form a branch. Therefore, this makes CHAID a very suitable algorithm for modelling natural hazards based on its multifaceted splitting in an optimum manner[35].

The splitting and merging parameters range between 0 and 1 and several parameters were used until the optimum parameters were set. After this process, the Chi-square statistic was applied by creating a tree structure which is initiated by the root node and further branched into the internal nodes, and afterwards the terminal nodes. The Chi-square establishes a binary decision that splits-up classes from other classes thereby creating a top-down structural tree till the terminal nodes are concluded. This creates a structure that relates the level of variable's influence on the tree's structure where some features are classified while others are rejected by the algorithm based on their relative importance.

4.5 Hybrid Modelling

The flood conditioning factors were all re-classified using the IOE weights (W_j). The database derived was then used as inputs in training the machine learning models (RF_{IOE}, SVM_{IOE} and DT_{IOE}). Ten-fold cross-validation was also performed on each of the hybrid machine learning models and the flood probability indexes were generated. To attest to the accuracy of the hybrid models, stand-alone ML models were also used in training the database in which all the influencing factors were all in a continuous data format and unclassified without the influence of the IOE weights deriving the flood probability indexes for each of the stand-alone machine learning models.

4.6 Creation of Flood Susceptibility Maps

The flood susceptibility models were used in deriving the flood susceptibility maps by splitting up flood probability index generated in continuous data format into pixels representing the susceptibility classes. The pixels are assigned a distinctive susceptibility index by calculating flood susceptibility indices for each pixel to ascertain the possibility of flood occurrences in the study area[22]. Each pixel obtained represents a value between 0 and 1, where 0 denotes no potential of flood susceptibility and 1 denotes a high potential of flood susceptibility.

The probability index was further classified using the quantile technique to produce the final susceptibility maps. The quantile technique was adopted in classifying the flood susceptibility index due to its suitability in grouping an equal number of pixels in the same classes without tampering with the values[34]. Quantile technique is a popular and efficient classification technique based on its reliable performance in hazard modelling[29]. Subsequently, the flood susceptibility maps were classified into five classes namely: low, very low, moderate, high, and very high based on literature[22].

4.7 Results Validation and Model's Performance Assessment

The performance assessment and prediction capability of the flood models are evaluated using both the training and the testing datasets through the ROC curve and statistical metrices (Step F). Both evaluation measures are described below.

4.7.1 Model Evaluation using the ROC Curve

The Receiver Operating Curve (ROC) curve is a popular, comprehensive evaluation tool used in hazard modelling studies[20], [52]. The ROC curve is conventionally utilized to assess the performance of the ML models and its efficiency relies on the ranking model's performance in an organized manner and attractive visualization[35], [48]. The statistical indicator of the ROC curve is represented by the Area Under Curve (AUC). Through the AUC, the accuracy and the performance of the ML models are quantified using various thresholds of the probability.

The curve is created by plotting 'sensitivity' on the Y-axis against '1-specificity' on the X-axis. The AUC ranges between 0 and 1 where 1 indicates that the observed and simulation data are in a perfect spatial agreement[1]. Therefore, the closer the value is to 1 determines the efficiency and the precision of the model. Consequently, an AUC value of <0.6 depicts weak accuracy, 0.6 – 0.7 indicates moderate accuracy while 0.7 – 0.8 depicts good accuracy while >0.8 indicates an almost perfect accuracy[66]. To plot the AUC curve and derive the statistical indices is dependent on the following parameters namely: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), (P) and (N). The AUC is derived as thus:

$$\boldsymbol{AUC} = \frac{\sum \boldsymbol{TP} + \sum \boldsymbol{TN}}{(\boldsymbol{P}+\boldsymbol{N})} \tag{21}$$

Where TP denotes the number of pixels correctly classified, TN denotes the number of pixels correctly classified as non-flood pixels, P is the total number of flood pixels, and N represents the total number of non-flood pixels.

Considering this study's context, success rate and prediction rate were constructed. The success rate represents a ROC curve plot type that describes how the flood probability index segregates the flood locations across the susceptibility zones and highlights the model fitting rate to the training dataset[4]. On the other hand, the prediction rate reveals the performance of the models in predicting locations prone to flooding and indicates how efficient the model is in predicting floods[3]. The Success Rate is constructed based on the training dataset which does not describe the efficiency of the model and the testing dataset was used in constructing the prediction rate through the comparison of the testing dataset to the flood susceptibility maps.

4.7.2 Statistical Metrics

Statistical measures were implemented in this study to augment the AUC curve in having a detailed statistical analysis of the model's predictive capabilities and to check the statistical significance of the models[3]. The set of statistical metrics considered in this study were Sensitivity (Recall), Accuracy and Specificity, Positive Predictive Rate (Precision), and Negative predictive Rate (NPV).

Sensitivity is a statistical index that measures the proportion of flood pixels correctly classified as flood pixels. Specificity index indicates and measures the proportion of non-flood pixels correctly classified as non-flood pixels. The accuracy index measures the rate of difference between the flood and non-flood pixels[4]. Each of the statistical indicators is defined as[52]:

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} * \mathbf{100} \quad (22)$$

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{FP+TN}} * \mathbf{100} \quad (23)$$

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP`+FP+TN+FN}} * \mathbf{100} \quad (24)$$

$$\boldsymbol{PPV} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} * \mathbf{100} \quad (25)$$

$$\boldsymbol{NPV} = \frac{\boldsymbol{TN}}{\boldsymbol{TN+FN}} * \mathbf{100} \quad (26)$$

Where FP equals the number of pixels incorrectly classified as flood events, and FN equals the number of pixels incorrectly classified as non-flood. Thus, the higher the TP and the lower the FP indicates the efficiency of the model.

4.8 Software and Device Specifications

The data derivation and pre-processing were implemented using the software ArcGIS Pro 2.7.0 from ESRI and ERDAS IMAGINE 2020. Afterwards, the IOE modelling was developed using ArcGIS Pro 2.7.0 and Microsoft Excel through which each coefficient values were derived.

WEKA 3.8.4 was used in performing the feature selection process while the multi-collinearity statistical analysis was derived through the SPSS statistics 26 software from IBM. The machine learning model training and classification was implemented using WEKA 3.8.4 which contains the required packages for conducting machine learning analysis as it includes the Decision Tree, Support Vector Machine and Random Forest classification algorithms and other packages which were used in performing the complete model training process. SPSS statistics 26 was also used in deriving the Area under Curve (AUC) and other performance metrics in validating the model.

The hardware (computer) utilized has an installed 8GB RAM and a processor Intel(R) Core i5-6300U CPU @2.40GHZ 2.50GHz.

5 RESULTS

This chapter introduces the outputs attained from the methodology implemented in chapter 4. It should be noted that only selected tables and figures where appropriate were presented while other results are attached in Annexes.

5.1 Predictiveness of Flood Influencing Factors (Feature Engineering)

5.1.1 Feature Selection

The predictive capability of each flood influencing factors on flood occurrences is germane to performing flood susceptibility. Therefore, the linear support vector machine algorithm was employed in performing the feature selection process and the Average Merit (AM) values were attained (Figure 6). The AM values describe the strength of each influencing factor in predicting flood occurrence and the value ranges between 0 and 1. Consequently, the distance from river obtained the highest AM (0.850) and followed by the population which obtained the average merit of 0.642, Distance from Road (0.583) while NDVI had the lowest average merit of 0.089. Based on the results attained, all the influencing factors achieved values greater than zero. Therefore, all the factors were considered in the flood susceptibility modelling as each factor tend to have a certain influence on flood occurrence. Furthermore, to ensure an outright assessment of flood susceptibility within the region, it is of high necessity to give full consideration for each of the influencing factors no matter how low the influence might be.

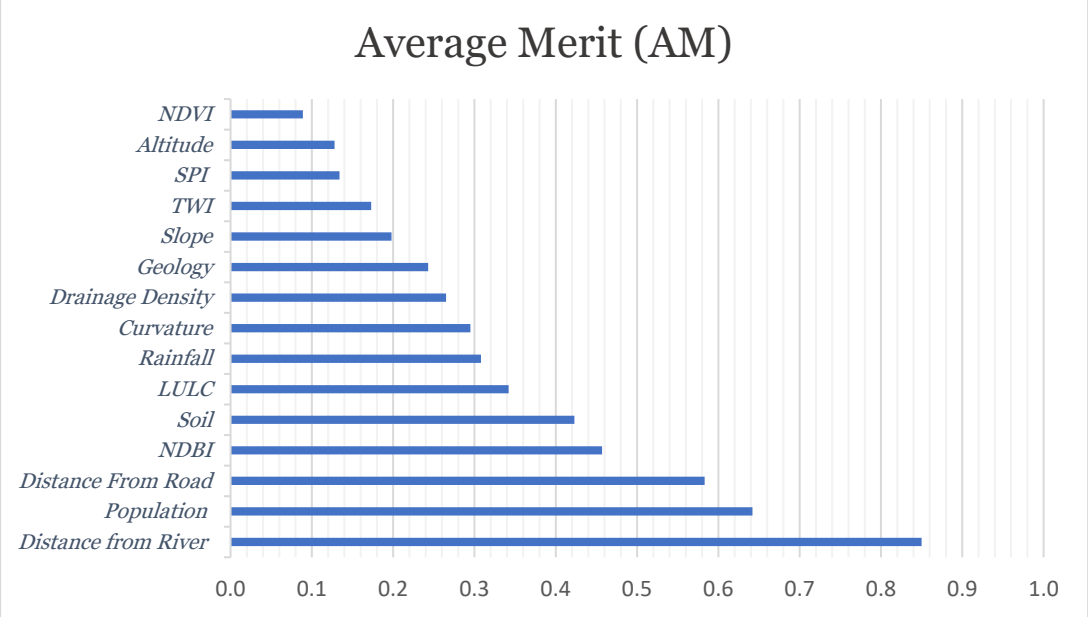


Figure 6: Factor's predictive ability (average merit) result

5.1.2 Multi-collinearity Analysis

Tolerance (TOL) and Variance Inflation Factor (VIF) was utilized in performing this process using the threshold of TOL > 0.1 and VIF < 10 critical values based on existing literature (Table 2). Geology has the lowest tolerance (0.115) and the highest VIF (8.714). However, even as it seems to have almost the same influence as soil with the TOI (0.133) and VIF (7.504), both influencing factors were considered as previous studies have proved the influence of geology on the occurrence of floods. Furthermore, all the influencing factors considered met the threshold laid down as all factors exceed the theoretical critical values for any evidence of multi-collinearity and were all therefore utilized in the modelling process.

Factor	Tolerance	VIF
Rainfall	0.500	2.001
Altitude	0.569	1.758
Curvature	0.736	1.359
Slope	0.458	2.183
TWI	0.430	2.326
SPI	0.685	1.460
Drainage Density	0.883	1.133
Distance from Road	0.439	2.278
LULC	0.443	2.258
Soil	0.133	7.504
Geology	0.115	8.174
Population	0.495	2.021
Distance from River	0.350	2.857
NDVI	0.359	2.782
NDBI	0.413	2.419

Table 2 : Multi-collinearity Analysis

5.2 Flood Modelling Algorithms

5.2.1 Index of Entropy Flood Modelling

IOE modelling was utilized in this study as bivariate and multivariate statistical modelling. This is to ensure the weights attributed to each class of an influencing factor and the overall weight of each factor on flood occurrence was attained (Table 3). To perform IOE modelling required calculating the FR values of the classes of each factor. However, it should be

mentioned that ratios > 1 signifies a high probability of flood occurrence while ratios < 1 signifies a low probability of flood occurrence. Population class between 21970 – 73351 attained the highest value of FR attaining 4.01. This was followed by grassland class of the LULC factor which attained a value of 2.56. It should be mentioned that 22 classes attained FR value equal to zero. The FR values were used to attain the probability density values (P_{ij}) which measures the probability of flood occurrence representing the weight of influence in each class of influencing factors. Consequently, distance from river class between 0 – 2986.4m attained the highest value of 0.98, followed by distance from road class 0 - 440.3m attaining a value of 0.85. This was then followed by the subrecent alluvium and coastal plain sands class of the geology factor which attained a value of 0.43. It should be noted that as in the case of FR values which attained the values of 0 also resulted in the probability density of the classes' values equaling to 0.

Subsequently, the weights of the flood influencing factors which ranges from 0 to 1 were derived after completing the modelling (Table 3). Distance from river achieved the highest weight with a value of 0.52 followed by the distance from road with a value of 0.34, Altitude (0.26), SPI (0.20), NDBI (0.19), LULC (0.19), Geology (0.18), Population (0.17), Soil (0.17), Slope (0.12), NDVI (0.12), Drainage Density (0.07), Rainfall (0.071), TWI (0.04), Curvature (0.002). Thereafter, the susceptibility index FSI_{IoE} was derived through the multiplication of the influencing factor's weight (W_j) with the IOE coefficients earmarked to each class.

Thus, in deriving the flood susceptibility index, the final equation is as follows:

$$FSI_{IoE} = 0.51 * [Distance\ from\ Road] + 0.34 * [Distance\ from\ River] + 0.26 * [Altitude] + 0.20 * [SPI] + 0.19 * [NDBI] + 0.19 * [LULC] + 0.18 * [Geology] + 0.17 * [Population] + 0.17 * [Soil] + 0.12 * [NDVI] + 0.07 * [Drainage\ Density] + 0.07 * [Rainfall] + 0.04 * [TWI] + 0.002 * [Curvatue]$$

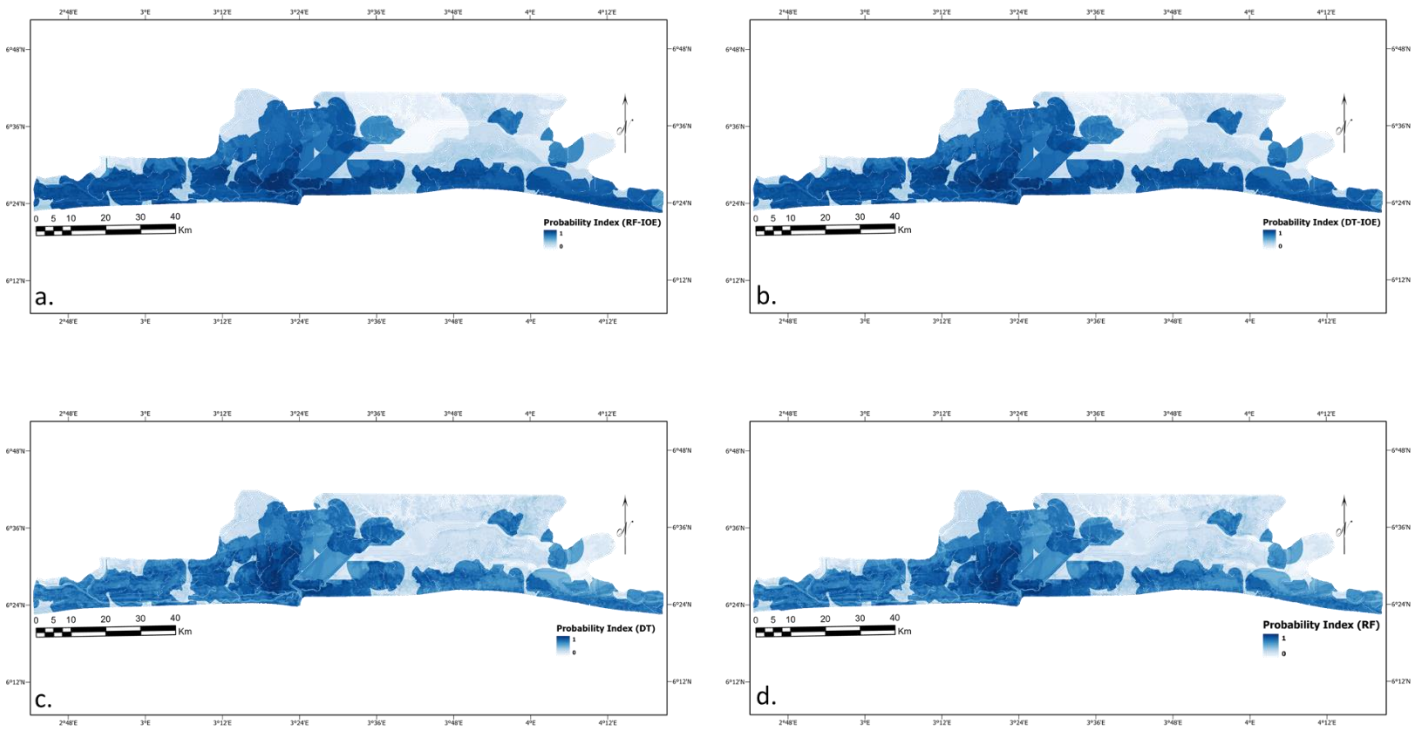
Factor	Class	Flood Pixels	Class Pixels	FR_{ij}	(P_{ij})	W_j
Altitude (m)	0 - 7	22	1621454	0.593882	0.158748	**0.258**
	8 - 16	54	1652308	1.43049	0.382377	
	20 - 36	19	538174	1.545301	0.413067	
	37 - 57	1	255416	0.17137	0.045808	
	58 - 109	0	134622	0	0	
Rainfall (mm)	0 - 9.6	7	948452	0.323004	0.089865	**0.071**
	9.7 - 10.7	48	1671607	1.2567	0.349637	
	10.8 - 12.0	34	1167270	1.274771	0.354664	
	12.1 - 13.7	7	414088	0.739827	0.205833	
Slope (Degree)	0 - 1.96	37	1921380	0.839859	0.21496	**0.124**
	1.97 - 5.15	42	1281650	1.429216	0.365805	
	5.16 - 9.78	12	645367	0.810947	0.20756	
	9.79 - 16.71	5	263677	0.827019	0.211674	
	16.72 - 45.32	0	74793	0	0	
Curvature	Convex	11	482290	0.998313	0.320926	**0.002**
	Flat	73	3255206	0.981582	0.315547	
	Concave	12	464478	1.130832	0.363527	
Soil	Aquic Paleudult	0	21150	0	0	**0.169**
	Orthic Tropudalf	0	19449	0	0	
	Oxic Rhodudalf	15	976453	0.671283	0.200306	
	Oxic Tropudalf	0	22323	0	0	
	Oxic Tropudalt	16	629038	1.111499	0.331663	
	Rhodic Paleudult	0	15477	0	0	
	Typic Tropaquent	0	16716	0	0	
	Typic Tropudult	2	535389	0.16324	0.04871	
	Typic Ustifluvent	63	1959055	1.40527	0.419322	
SPI	0 - 67626	96	4185641	1.000293	1	**0.2**
	676267 - 287411	0	743	0	0	
	287411 - 625452	0	315	0	0	
	625542 - 1115832	0	117	0	0	
	1115833 - 2155584	0	51	0	0	
Drainage Density	0 - 3896.93	52	2650331	0.858909	0.237492	**0.073**
	3897 - 11690.78	34	944863	1.575266	0.435568	
	11691 - 22948.57	9	493284	0.79871	0.220847	
	22949 - 55206.47	1	114092	0.383697	0.106094	
Distance from Road (m)	0 - 440.3	90	2138540	1.834335	0.847763	**0.337**
	441.4 - 9373.4	3	1242905	0.105205	0.048622	
	9374 - 19598.7	3	583243	0.224195	0.103615	
	19600 - 36214.5	0	219630	0	0	
Land use Land Cover	Irrigated Croplands	0	1040	0	0	**0.185**
	Residential Areas	17	728714	1.020756	0.09349	
	Rural Areas	0	11631	0	0	

	Broadleaved Evergereen Forest	15	871338	0.753242	0.068989	
	Semi-deciduous Woodland	8	403307	0.867929	0.079493	
	Broadleaved Deciduous Woodland	0	47792	0	0	
	Grassland	1	17097	2.559236	0.234398	
	Shrubland	6	239933	1.094187	0.100216	
	Hebaceous Vegetation	3	229139	0.572865	0.052468	
	Broadleaved Forest regularly flooded	0	44018	0	0	
	Broadleaved Forest	3	101851	1.288802	0.11804	
	Woody Vegetation	0	1557	0	0	
	Central Business District	38	665013	2.500251	0.228996	
	Water Bodies	5	838074	0.261046	0.023909	
Geology	Coastal Plain Sands	15	997603	0.657052	0.19921	**0.175**
	Coastal Plain Sands (Alfisols)	2	537657	0.162551	0.049284	
	Coastal Plain Sands (Ultisols)	16	651361	1.073406	0.325444	
	Deltaic Basin and Tidal Flats	0	15477	0	0	
	Ewekoro Formation (Upper Coal Measure)	0	16077	0	0	
	Recent Alluvium	0	17181	0	0	
	Sandstone (Abeokuta Formation)	0	639	0	0	
	Subrecent Alluvium and Coastal Plain Sands	63	1959055	1.40527	0.426062	
Population	0 - 486	52	3334648	0.682253	0.061751	**0.173**
	487 - 1053	13	336658	1.689454	0.152912	
	1054 - 6010	15	287226	2.284859	0.206802	
	6011 - 21969	9	165197	2.383594	0.215738	
	21970 - 73351	7	76405	4.008373	0.362797	
Distance from River(m)	0 - 2986.4	95	2469057	1.665271	0.97788	**0.513**
	3000.2 - 7700.3	1	1148992	0.037668	0.02212	
	7701.4 - 40071.7	0	536880	0	0	
NDVI	0 - 0.04160	1	298954	0.146425	0.031126	**0.122**
	0.04200 - 0.25140	38	873942	1.903355	0.4046	
	0.25150 - 0.43219	21	801044	1.147576	0.243943	
	0.43220 - 0.61302	13	656931	0.866248	0.18414	
	0.61400 - 0.84449	23	1571457	0.640684	0.136192	
NDBI	-1 - -0.4814	1	487770	0.089744	0.018046	**0.192**
	-0.4813 - -0.2320	10	1251171	0.349866	0.070354	
	-0.2319 - -0.0884	15	724405	0.906418	0.182269	
	-0.0883 - 0.0552	32	682483	2.05247	0.412726	
	-0.0553 - 1.0000	38	1056499	1.574466	0.316605	
TWI	0 - 6.8	31	1348395	1.00275	0.210479	**0.042**
	6.9 - 8.9	29	1154504	1.095597	0.229968	
	9.0 - 11.6	21	690239	1.326991	0.278538	
	11.7 - 14.3	11	480114	0.999302	0.209755	
	14.4 - 24.2	4	513908	0.339487	0.071259	

Table 3: Frequency ratio and Index of Entropy coefficients values distribution within flood influencing factors classes.

5.2.2 Support Vector Machine Flood Modelling

To perform the SVM flood modelling requires the tuning of parameters gamma (γ) and penalty (C) to train the SVM stand-alone and SVM-IOE hybrid model. This was achieved through the ten-fold cross-validation. In consideration of the SVM stand-alone model, $\gamma(0.11)$ and $C(1.4)$ were attained and used for optimizing the database and deriving the flood susceptibility index (Figure 7) while for the SVM-IOE model, $\gamma(0.19)$ and $C(1.6)$ were attained and used in generating the flood susceptibility index. The stand-alone SVM flood susceptibility map ranges from 0.152 to 0.836 and using the quantile technique, the map was classified into five classes of very low, low, moderate, high, and very high (Figure 8). The lowest class (0.152 – 0.369) occupies 31.73% of the total study area which indicates the low flood susceptibility areas of the watershed and the moderate class (0.495 – 0.635) occupies 20.47% of the watershed while the very high susceptibility class (0.680 – 0.836) which indicates areas highly prone to flooding occupies 22.81% of the watershed. On the other hand, SVM-IOE susceptibility map ranges from 0.196 to 1.374. SVM-IOE lowest susceptible class (0.196 – 0.450) occupies 21.56% of the total study area while the moderate flood susceptible class (0.810 – 1.032) occupies 18.86% of the watershed while the highest flood susceptible class (1.110 – 1.374) occupies 22.41% of the study area.



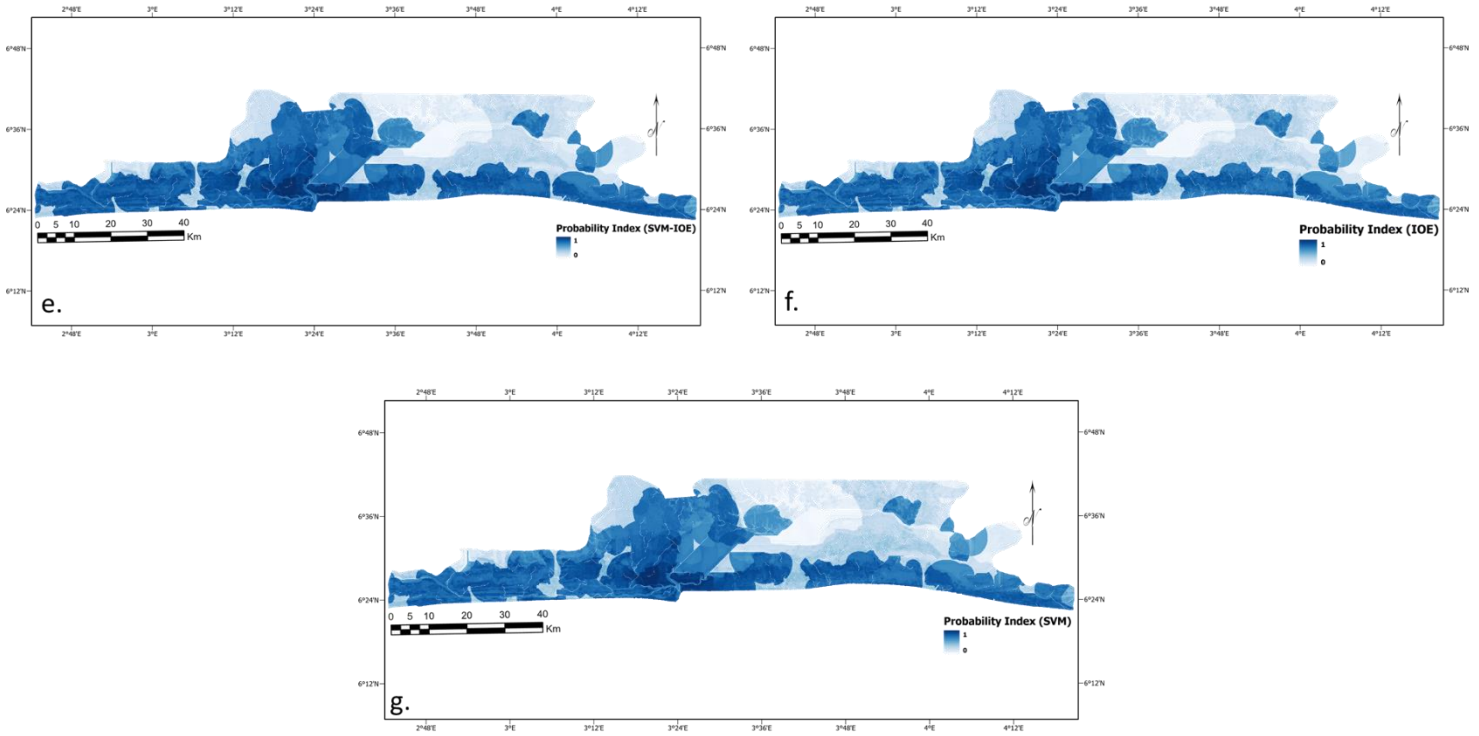


Figure 7: Flood probability index derived from: (a) RF-IOE, (b) DT-IOE, (c) Stand-alone DT, (d) Stand-alone RF, (e) SVM-IOE, (f) Stand-alone IOE, (g) Stand-alone SVM.

5.2.3 Random Forest Flood Modelling

The ten-fold cross-validation technique was also used in tuning the *mtry* and *ntree* parameters to train the random forest model. Thereafter, optimal parameters were set to be $ntree = 2000$ trees and $mtry = 10$. The out of bag error procedure was also implemented, which is based on the uniformity of the nodes and leaves within the RF model[58]. This is to ensure accuracy within the model as the model's accuracy decreases based on the exclusion of important variables. Based on these metrics, distance from road, distance from river, population, LULC and soil demonstrates high importance in the flood modelling.

The RF susceptibility map (Figure 8) derived ranges from 0.130 to 0.675. The lowest class (0.130 – 0.327) which signifies areas less susceptible to flood occupies 36.15% of the study area while about 15.60% of the watershed signifies areas moderately susceptible to flood within the watershed. The highest class (0.547 – 0.675) signifies areas highly susceptible and occupies 23.49% of the total study area. On the other hand, the optimal parameters achieved through the cross-validation for the RF-IOE was $ntree = 2500\ trees$ and $mtry = 11$ and the out of the bag procedure was also implemented for modelling the algorithm. Thus, the final RF-IOE susceptibility index ranges from 0.055 to 0.615. The lowest susceptible class (0.055 – 0.171) occupies 20.71% of the study area while the moderate

susceptible class (0.275 – 0.483) occupies 37.14% of the watershed having the highest percentage of the study area. The highest susceptible class (0.518 – 0.615) which signifies areas highly susceptible to flood occupies 17.32% of the study area.

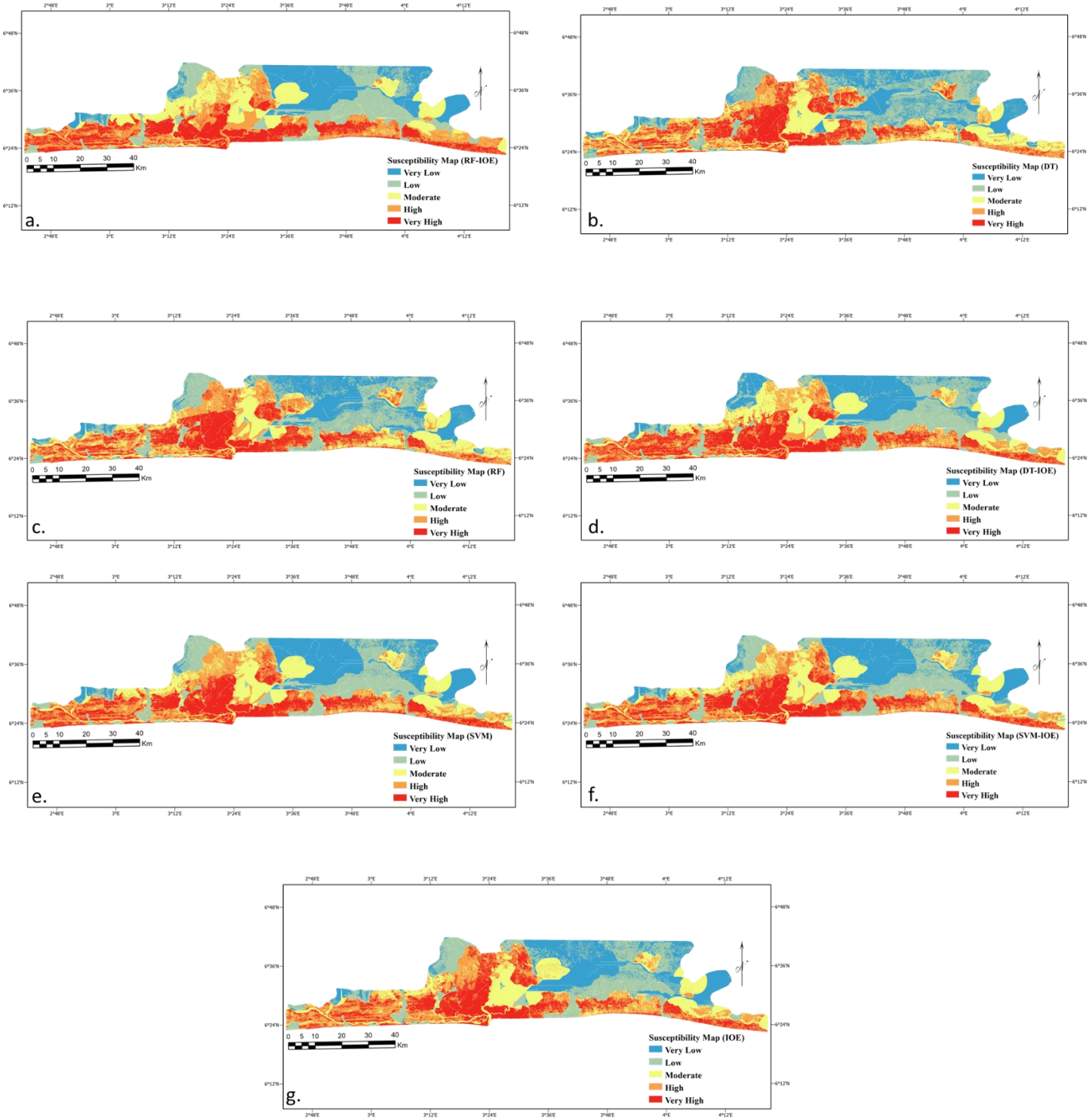


Figure 8: Flood susceptibility maps derived from (a) RF-IOE, (b) Stand-alone DT, (c) Stand-alone RF, (d) DT-IOE, (e) Stand-alone SVM, (f) SVM-IOE, (g) Stand-alone (IOE).

5.2.4 Decision Tree Flood Modelling

The splitting and merging parameters were set with the values of 0.8 and 0.001 for the stand-alone DT model which were arrived at after continuous trial and error to achieve the best accuracy. DT algorithm reduces parameters in performing its modelling as the most important variables are selected and used in generating the model. Therefore, in the stand-alone DT model, TWI, geology, altitude, NDVI and slope were rejected by the algorithm while the other 10 influencing factors were utilized for the modelling. In terms of the DT top-down structure, the influencing factors higher on the tree structure signifies a higher influence on flood occurrence. Therefore, distance from road, distance from river, drainage density, LULC and rainfall were first selected for splitting by the algorithm.

Thus, the DT tree generated contains 10 variables, 458 nodes and 108 leaves making it impossible to present as a tree in the document and each leaf describes a certain degree of flood potentiality. The final susceptibility index generated for the DT model ranges from 0.114 to 0.649. The lowest susceptible class (0.114 – 0.313) occupies 37.20% of the total study area while the moderate class (0.412 – 0.489) occupies 14.39% of the watershed. The highest susceptible class (0.529 – 0.649) indicating areas highly prone to floods occupies 22.43% of the study area. On the other hand, the splitting and merging parameters for the DT-IOE model was set to 0.7 and 0.005, respectively. NDVI, TWI, slope, and altitude were rejected by the hybrid model.

Therefore, 11 influencing factors were utilized for the modelling and the final tree contains 648 nodes and 189 leaves. The flood susceptibility index derived ranges from 0.057 to 0.600. The lowest class of susceptibility (0.057 – 0.168) occupies 18.50% of the study area while the third class (0.265 – 0.467) which signifies areas moderately susceptible to flood occupies 37.20% of the study area having the highest percentage of the study area. The highest susceptible class (0.503 -0.600) which indicates high flood susceptibility occupies 17.86% of the study area.

5.3 Accuracy Assessment and Validation of Flood Models

The seven flood susceptibility maps were evaluated by AUC technique and other statistical metrics namely Sensitivity, Specificity, Accuracy, Positive Predictive Value (PPV), and Negative Predictive Value (NPV). This was to test the accuracy of the models and the higher the AUC values the better the model in terms of success rate and prediction rate. The DT and RF stand-alone model has the highest success rate of 0.900 while the IOE model and the SVM model both had the success rate of 0.899 (Figure 9).

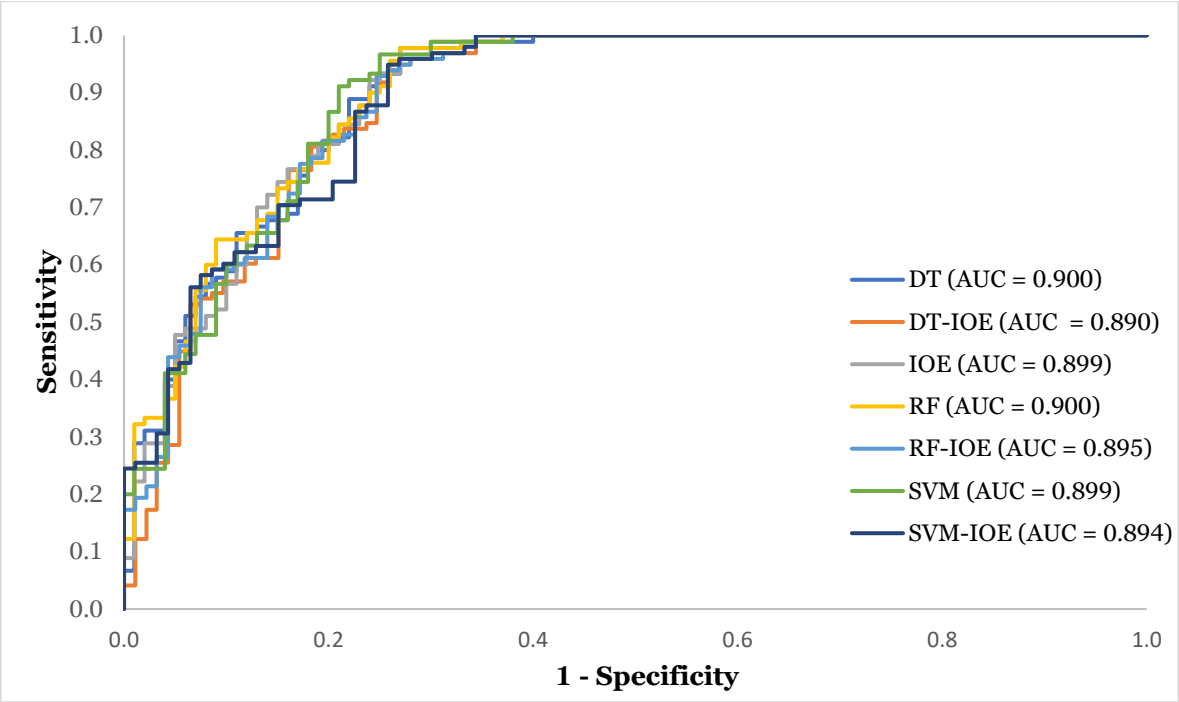


Figure 9: Area under Curve (AUC) showing the Success Rate

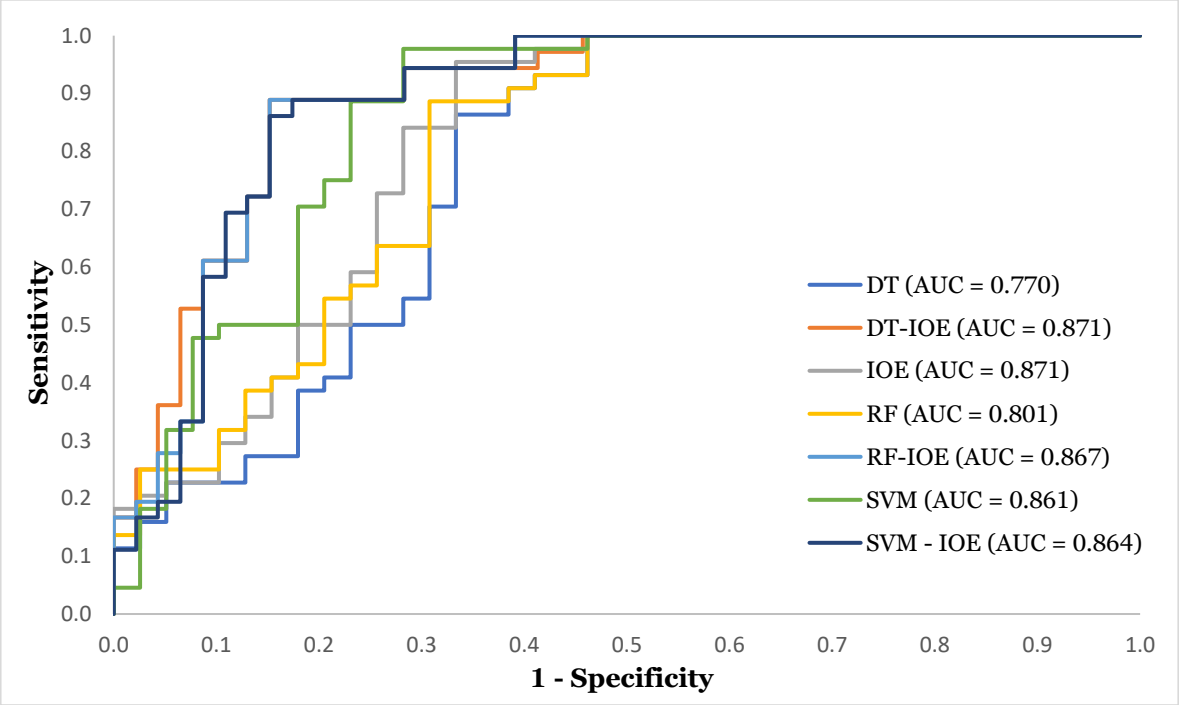


Figure 10: Area under Curve (AUC) showing the Prediction Rate

This was then followed by the RF-IOE (AUC = 0.895), SVM-IOE (AUC = 0.894), and DT-IOE (AUC = 0.890) had the lowest performance.

In consideration of the hybrid models, DT-IOE model produced the highest prediction rate of 0.871 followed by the RF-IOE model with a prediction rate of 0.867 while SVM-IOE had a prediction rate of 0.864 (Figure 10). On the other hand, the DT stand-alone model produced the lowest prediction rate of 0.770, followed by the RF stand-alone model with a prediction rate of 0.801 while the stand-alone SVM model had the highest prediction of 0.861. Percentagewise, using the hybrid technique improved the prediction rate as the DT-IOE is 10% higher than the stand-alone model, and the RF-IOE model is 7% higher than RF stand-alone model while the SVM-IOE slightly performed than the SVM stand-alone model in predicting areas susceptible to flood.

Classifier	Sensitivity	Specificity	Accuracy	PPV	NPV
DT-IOE	83.2	77.6	80.3	78.2	82.6
DT	67.4	69.2	68.3	70.7	65.9
IOE	74.4	72.1	73.2	70.7	75.6
RF	67.4	69.2	68.3	70.7	65.9
RF-IOE	79.4	80.6	80.3	81.1	78.9
SVM	78.0	69.8	73.8	71.1	76.9
SVM-IOE	82.4	84.1	78.3	77.0	80.0

Table 4: Performance metrics of Classifiers

The authenticity and reliability of the results were further checked with other statistical metrics (Table 4). Overall, the DT-IOE model had the highest performance assessing the accuracy of the classifiers based on the validation dataset with a specificity (83.2%), and this was followed by SVM-IOE with the sensitivity of 82.4%, RF-IOE (79.4%), SVM (78%), IOE (74.4%), RF (67.4%) and DT (67.4%). Based on the accuracy, the RF-IOE and DT-IOE had the overall highest performance with an accuracy of 80.3%, followed by the SVM-IOE (78.3%), SVM (73.8%), IOE (73.2%), RF (68.3%) and DT (68.3%).

6 DISCUSSION

The major focus of this research is to develop and utilize machine learning-based flood susceptibility models in deriving flood susceptible maps considering a diverse range of factors relative to the study area and to identify the impact of the factors on flood occurrence in the study area. This was performed taking account of the human-induced factors and other natural-caused factors acknowledged in previous studies and proven to have a certain influence on flood occurrence in the study area. Moreover, feature engineering was performed to investigate the predictive ability, significance, and interrelationship among the influencing factors before the main modelling and all the factors had a certain influence and were therefore utilized for the modelling. Furthermore, feature engineering was performed as there are no existing works related to flood susceptibility in the area which necessitates the identification of the factors prevalent in the region and their significance.

Also, it is worthy to note the importance of remote sensing which facilitated the derivation of most of the influencing factors, verification of the past flood events and generally plays a huge role in assessing potential areas susceptible to natural hazards. Subsequently, in achieving accurate results and high prediction based on the novelty of FSM in the region, novel hybrid models of efficient ML algorithms namely DT, SVM and RF integrated with IOE, and each model as a stand-alone model was implemented in deriving flood susceptibility maps for the study area (Figure 8). The ML models were implemented as an effectual flood risk assessment is essential due to the increasing incessant flooding in the region, the complex nature of the flood influencing variables resulting in its occurrence and its long devastating effects on the region.

Based on the modelling's novelty in the area, the study utilized all the fifteen influencing variables considering a certain degree of influence each variable has on the spatial distribution of flood occurrence in the region. The utilization of the IOE statistical model provided the platform in assessing the influence of each class of each influencing factor and the overall weight of each factor on flood occurrence through bivariate and multivariate statistical analysis (Table 3). The acquired weights were then used in reclassifying the factors. IOE was further utilized based on its superiority to other BSA/MSA models as it does not presume a linear model and make no assumptions with regards to the distribution of variables[53].

According to Costache et al. (2020) and Tehrany et al.(2019), adopting this hybridization approach of machine learning and statistical techniques approach increases the accuracy of the classification algorithms significantly which was further proven by the outcome of this

study[35]. Furthermore, the influence of each class of each factor's influence on the occurrence of floods is identified and given uttermost consideration.

Regarding the first research question, distance from river network has the highest influence on flood occurrence in the area. This confirms that areas highly susceptible to floods are very close to the river network, Furthermore, the lowest class (0 – 2986.4m) of the distance from river has the highest influence (0.98) which indicates areas within 3000m to the water bodies are highly susceptible to flooding. This is followed by distance from road which had the second overall influence on flood occurrence within the region. This was more significant in the factors first class (0 – 440.3m) which signifies the confluence of roads within a 440m radius highly influences the occurrence of floods.

Furthermore, altitude does have a significant influence on the occurrence of floods in the area having the highest third overall influence. Water flows from high altitudes to lower ones and therefore upsurges the occurrence of floods in the lower areas. The geologic composition of the region influences flooding as the subrecent alluvium and coastal plain lands which comprises 85% of the Lagos landform is composed of fossils, sedimentary rocks which reduces water permeability, allowing water inundation, and exacerbates flooding. This reveals the impact of geology on flooding within the region.

Consequently, the overall results reveal that the factors majorly distance from river, distance from road, NDBI, altitude, soil, LULC, rainfall and SPI play a critical and huge role in the spatial distribution of flood occurrences in the study area. On the other hand, slope has a low influence on flooding in the study area as the slope reflects a downward trend thereby having a huge percentage of flat curvature which increases runoff speed and reduced water percolation into the soil. However, this affected the SPI which describes the erosion capacity to be greatly increased thereby having a very moderate impact on flood occurrence. Also, the NDVI which describes the vegetation density has a low impact on flood occurrence in the region due to the low vegetation density in the urban and residential areas. Curvature had the lowest value signifying the least influence on the occurrence of floods.

Based on previous studies, curvature often generates low values which have, in turn, led to various contrasting assertions on the role of curvature in FSM[20]. Figure 9. 1(Annex) details the relative distribution of flood pixels within the classes of each flood influencing factor which describes the flood potential observed in each class. Based on the flood susceptibility maps derived, areas highly prone to flooding occupied approximately 21% of the study area, which are around the lagoon, also in the plain region near river networks and majorly in the residential areas and the central business district of the city. This reveals the impact of urbanization on flood occurrences in the study area and confirms the influence of

NDBI and LULC on the occurrence of floods in the region. This can be attributed to Bahram Choubin's study that revealed 181,000 people inhabiting residential and urban areas around rivers in the world are consistently affected by flooding[45].

Regarding the second research question based on accuracy assessment and predictive performance of the models, all the models performed relatively well with slight differences in terms of the success rate which describes the model fitting rate to the training dataset. The DT and RF models had the highest performance followed by SVM and IOE. however, hybrid models (DT-IOE, SVM-IOE, and RF-IOE) attained the lowest performance which could be attributed to quantile classification technique implemented in classifying each of the influencing factors in performing the IOE modelling though quite insignificant based on the success rates achieved.

On the other hand, considering the prediction rate, the DT-IOE had the highest overall prediction performance over the six other models outperforming the SVM-IOE and the RF-IOE models. This signifies that the DT-IOE susceptibility map had the highest accuracy in predicting flood locations prone to flooding. Also, this proves the rejection or addition of additional influencing factors does not necessarily increase the accuracy of the models as factors with similar influence may be rejected by the algorithm and does not suggest the factor's significance invalid[35].

Tehrany et al. (2019) also attained similar findings where DT model performed slightly more than SVM model. This relates to the SVM's ability in handling multi-linear data with high precision and low error rates. Fotovatikhah et al. (2018) explored over a hundred articles on floods and attested to this fact[43]. However, SVM algorithm lacks the ability in evaluating the significance of variables utilized. Consequently, based on literature, past flooding studies often utilizes SVM with various statistical models and integration with other machine learning models in addressing the importance of variables utilized[4], [35]. However, DT stand-alone model attained the lowest prediction performance.

Furthermore, the advantage of utilizing hybrid machine learning models is further proven as the DT-IOE obtained a higher accuracy than the DT stand-alone model. Additionally, it should be mentioned that the utilization of hybrid models quickened the modelling process as the processing time remarkably reduced due to the pre-processing of the flood influencing factors as one of the major drawbacks of the ML models utilized is the required time for analysis.

On the other hand, to ensure statistical significance and overall efficiency, statistical metrics namely Sensitivity, Specificity, Accuracy, NPV and PPV were utilized in also assessing the performance of the models. The DT-IOE performed best slightly than the RF-IOE and the

SVM-IOE in terms of specificity of over 80%. The RF had the lowest performance of approximately 70% sensitivity which suggest a good indiscriminatory performance of all the models utilized for the study.

Also, each model produced partly varied different susceptibility patterns with relation to the susceptibility maps (Figure 8) generated even with similar statistical performances. This could be relatively attributed to the selection procedure and the utilization technique of the variables implemented by each ML algorithm. The RF permutes each variable randomly and DT (CHAID) selects the variables hierarchically based on their relative importance on flood occurrence while SVM incorporates all the variables and sets up an optimal plane to distinguish the classes of the variables based on flood and non-flood origin. Thus, the variation of susceptibility patterns based on selection and combinations of set of factors by each ML algorithm.

However, it should be mentioned that selecting the best model for FSM is quite challenging, even though hybridization of models is powerful, variations exist depending on the region which could induce uncertainty in spatial prediction. Therefore, changing input data based on future conditions could alter the model's accuracy[19].

6.1 Limitations and Recommendations

There were some remarkable limitations encountered in this study. Foremostly, there was data paucity (spatial and temporal) in terms of flood influencing variables, flood inventory data, and resolution of imageries acquired for the study. It is ascertained that the prediction abilities of the factors will increase if the factors are derived from higher resolution imageries. Also, regarding the temporal dimension of flood occurrence in the region, which is fully based on rainfall that initiates flooding, there is a limitation based on the availability of rainfall data from rainfall stations which could reveal the unceasing influence of rainfall on flooding in the region. However, rainfall's influence was quite significant in the study based on the data utilized which relates that more reliable data will further reveal the increasing influence of rainfall in the study area over time.

Furthermore, acquiring more accurate and detailed inventory data is very fundamental to the entire process and would help in enriching and optimizing the model's parameters for the study area and more accurate models can be attained. Even though multiple iterations of random points for the inventory map were performed, there was no increase in the precision of models which addresses the stability of the results attained.

However, it is needed to acquire more inventory points where one inventory dataset (map) can be used for training and another inventory dataset for testing and multiple interactions can be performed with their outcomes compared. This will help to attain a more significant impact on the temporal dimension of flood occurrence in the region. In essence, there is a huge absence of a comprehensive spatial data infrastructure (SDI) in the region which is essential in bringing new insights into the flood susceptibility domain.

Also, based on feasibility studies that have been carried out in the region, wastes generation and disposal is a key factor in the occurrence of floods due to the blockage of water channels and decrease in the level of water percolation[12]. Therefore, more investigation is needed towards the exploration of this factor by acquiring tons of waste generated per block radius in each flood susceptibility zone.

In conclusion, performing accurate flood susceptibility mapping requires updated and accurate flood historical data, high resolution derived flood influencing variables and a powerful modelling algorithm to achieve highly sustainable results.

7 CONCLUSION

Flood Susceptibility Modelling (FSM) is one of the most popular research areas in natural hazard studies. It is a significant domain where accuracy and time are essential in mitigating and preventing flood occurrences. This study was implemented to investigate drivers of flooding and identify areas prone to flooding in the West Africa region using Lagos as a case study. This was achieved by implementing a novel hybrid and stand-alone machine learning algorithms specifically DT-IOE, SVM-IOE, RF-IOE, DT, SVM and RF to train the geospatial database composed of 15 influencing factors and 139 flood locations and 139 non-flood locations. The hybrid models were created to enhance the accuracy of the stand-alone models.

Thereafter, the models were then validated using the AUC and statistical metrics to check the statistical significance and the overall efficiency of the model. Based on the AUC through the evaluation of the success rate and the prediction rate, the DT-IOE (AUC = 0.899) achieved better goodness of fit to the training dataset and the highest prediction accuracy (AUC = 0.871). Percentagewise, the DT-IOE was approximately 10% higher than the DT stand-alone model and the RF-IOE was 7% higher than RF stand-alone model which is a significant improvement based on the model's prediction accuracy. On the other hand, checking for the statistical significance of the models and overall efficiency in terms of Accuracy, Sensitivity and Specificity, DT-IOE had the best performance based on the validation dataset which concludes DT-IOE as the most appropriate ML algorithm when natural-caused and human-induced factors are concerned with regards to the study area.

The results revealed that human-induced factors play a huge role in the occurrence of floods such as distance from road, NDBI and population density while natural-caused factors such as distance from river, LULC, geology proved to be very significant drivers of flood occurrences in the region. Also, the performance of feature selection process was implemented to identify the most significant factors before performing the main modelling. As a novel-based study in the region, the susceptibility maps generated would assist the urban planners to prevent the increasing urbanization in the identified susceptible regions thereby mitigating flood impact. Also, more inclined floodplains management approaches and refined policies can be developed. Over time, it is realized that LULC, SPI and rainfall as spatio-temporal factors influences the occurrence of flood and significant attention should be given to the factors as a change in LULC over time determines a significant impact of SPI and rainfall in the region.

In summary, the contribution of this research is emphasized as follows:

1. The proposed integration of IOE with DT, SVM and RF is a powerful modelling tool in the classification and identification of flood locations prone to flood risk.
2. The adoption of feature engineering technique is a considerable approach before the performance of FSM to identify the significance of the flood drivers as these factor's influence varies with time as to the adoption of user-defined factors from previous research.
3. Human-Induced factors should be given full consideration in any region as significant drivers of flood occurrences.

Finally, it should be mentioned that adopting the utilization of machine learning and geospatial technology is very efficient in performing FSM based on time, costs, and accuracy without expert judgement in the modelling. Also, the study gave insights into the carrying out of FSM within West Africa as the results attained is relevant to the national and local governments of flood-prone countries within the region which proves FSM can be carried out successfully in the region.

8 BIBLIOGRAPHIC REFERENCES

[1] Y. Wang, Z. Fang, H. Hong, and L. Peng, "Flood susceptibility mapping using convolutional neural network frameworks," *J. Hydrol.*, vol. 582, no. March, p. 124482, 2020, doi: 10.1016/j.jhydrol.2019.124482.

[2] R. Mind'je *et al.*, "Flood susceptibility modeling and hazard perception in Rwanda," *Int. J. Disaster Risk Reduct.*, vol. 38, no. April 2018, p. 101211, 2019, doi: 10.1016/j.ijdrr.2019.101211.

[3] W. Chen *et al.*, "Modeling flood susceptibility using data-driven approaches of naïve Bayes tree, alternating decision tree, and random forest methods," *Sci. Total Environ.*, vol. 701, 2020, doi: 10.1016/j.scitotenv.2019.134979.

[4] R. Costache *et al.*, "Spatial predicting of flood potential areas using novel hybridizations of fuzzy decision-making, bivariate statistics, and machine learning," *J. Hydrol.*, vol. 585, no. December 2019, p. 124808, 2020, doi: 10.1016/j.jhydrol.2020.124808.

[5] I. E. Olorunfemi, A. A. Komolafe, J. T. Fasinmirin, A. A. Olufayo, and S. O. Akande, "A GIS-based assessment of the potential soil erosion and flood hazard zones in Ekiti State, Southwestern Nigeria using integrated RUSLE and HAND models," *Catena*, vol. 194, no. January, p. 104725, 2020, doi: 10.1016/j.catena.2020.104725.

[6] P. T. Padi, G. Di Baldassarre, and A. Castellarin, "Floodplain management in Africa: Large scale analysis of flood data," *Phys. Chem. Earth*, vol. 36, no. 7–8, pp. 292–298, 2011, doi: 10.1016/j.pce.2011.02.002.

[7] I. Ajibade, G. McBean, and R. Bezner-Kerr, "Urban flooding in Lagos, Nigeria: Patterns of vulnerability and resilience among women," *Glob. Environ. Chang.*, vol. 23, no. 6, pp. 1714–1725, 2013, doi: 10.1016/j.gloenvcha.2013.08.009.

[8] J. Ntajal, B. L. Lamptey, I. B. Mahamadou, and B. K. Nyarko, "Flood disaster risk mapping in the Lower Mono River Basin in Togo, West Africa," *Int. J. Disaster Risk Reduct.*, vol. 23, no. October 2016, pp. 93–103, 2017, doi: 10.1016/j.ijdrr.2017.03.015.

[9] I. Douglas, "Flooding in African cities, scales of causes, teleconnections, risks, vulnerability and impacts," *Int. J. Disaster Risk Reduct.*, vol. 26, no. September, pp. 34–42, 2017, doi: 10.1016/j.ijdrr.2017.09.024.

[10] C. C. Olanrewaju, M. Chitakira, O. A. Olanrewaju, and E. Louw, "Impacts of flood disasters in Nigeria: A critical evaluation of health implications and management,"

Jamba J. Disaster Risk Stud., vol. 11, no. 1, pp. 1–9, 2019, doi: 10.4102/jamba.v11i1.557.

[11] E. I. D. Database, "EM-DAT." https://www.emdat.be/ (accessed Oct. 13, 2020).

[12] A. O. Israel, "Nature, the built environment and perennial flooding in Lagos, Nigeria: The 2012 flood as a case study," *Urban Clim.*, vol. 21, pp. 218–231, 2017, doi: 10.1016/j.uclim.2017.06.009.

[13] U. C. Nkwunonwo, M. Whitworth, and B. Baily, "A review of the current status of flood modelling for urban flood risk management in the developing countries," *Sci. African*, vol. 7, p. e00269, 2020, doi: 10.1016/j.sciaf.2020.e00269.

[14] K. Chapi *et al.*, "A novel hybrid artificial intelligence approach for flood susceptibility assessment," *Environ. Model. Softw.*, vol. 95, pp. 229–245, 2017, doi: 10.1016/j.envsoft.2017.06.012.

[15] N. U. Whitworth Malcolm, "Flooding and Flood Risk Reduction in Nigeria: Cardinal Gaps," *J. Geogr. Nat. Disasters*, vol. 05, no. 01, pp. 1–12, 2015, doi: 10.4172/2167-0587.1000136.

[16] S. V. Razavi Termeh, A. Kornejady, H. R. Pourghasemi, and S. Keesstra, "Flood susceptibility mapping using novel ensembles of adaptive neuro fuzzy inference system and metaheuristic algorithms," *Sci. Total Environ.*, vol. 615, pp. 438–451, 2018, doi: 10.1016/j.scitotenv.2017.09.262.

[17] C. Ugonna, "A Review of Flooding and Flood Risk Reduction in Nigeria," *Glob. J. Human-Social Sci.*, vol. 16, no. 2, 2016.

[18] S. Janizadeh *et al.*, "Prediction success of machine learning methods for flash flood susceptibility mapping in the Tafresh watershed, Iran," *Sustain.*, vol. 11, no. 19, 2019, doi: 10.3390/su11195426.

[19] H. Shafizadeh-Moghadam, R. Valavi, H. Shahabi, K. Chapi, and A. Shirzadi, "Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping," *J. Environ. Manage.*, vol. 217, pp. 1–11, 2018, doi: 10.1016/j.jenvman.2018.03.089.

[20] H. Hong, P. Tsangaratos, I. Ilia, J. Liu, A. X. Zhu, and W. Chen, "Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China," *Sci. Total Environ.*, vol. 625, pp. 575–588, 2018, doi: 10.1016/j.scitotenv.2017.12.256.

[21] M. S. Tehrany, B. Pradhan, and M. N. Jebur, "Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS," *J. Hydrol.*, vol. 504, pp. 69–79, 2013, doi:

10.1016/j.jhydrol.2013.09.034.

[22] M. S. Tehrany, L. Kumar, and F. Shabani, “A novel GIS-based ensemble technique for flood susceptibility mapping using evidential belief function and support vector machine: Brisbane, Australia,” *PeerJ*, vol. 2019, no. 10, 2019, doi: 10.7717/peerj.7653.

[23] P. Times, “What Lagos govt, residents must do to check flooding in Lekki,” 2017. https://www.premiumtimesng.com/regional/ssouth-west/241903-lagos-govt-residents-must-check-flooding-lekki-nimet.html (accessed Jan. 12, 2021).

[24] TheGuardian, “Lagos and the coming rains,” 2020. https://guardian.ng/opinion/lagos-and-the-coming-rains/ (accessed Jan. 12, 2021).

[25] S. Samanta, D. K. Pal, and B. Palsamanta, “Flood susceptibility analysis through remote sensing, GIS and frequency ratio model,” *Appl. Water Sci.*, vol. 8, no. 2, pp. 1–14, 2018, doi: 10.1007/s13201-018-0710-1.

[26] M. Shafapour Tehrany, L. Kumar, M. Neamah Jebur, and F. Shabani, “Evaluating the application of the statistical index method in flood susceptibility mapping and its comparison with frequency ratio and logistic regression methods,” *Geomatics, Nat. Hazards Risk*, vol. 10, no. 1, pp. 79–101, 2019, doi: 10.1080/19475705.2018.1506509.

[27] O. of D. R. R. (UNDRR), “Global Assessment Report on Disaster Risk Reduction,” 2019. https://digitallibrary.un.org/record/3825375?ln=en#:~:text=The Global Assessment Report on,efforts to reduce disaster risk.&text=The GAR aims to focus,economic support for risk reduction.

[28] U. C. Nkwunonwo, M. Whitworth, and B. Baily, “Review article: A review and critical analysis of the efforts towards urban flood risk management in the Lagos region of Nigeria,” *Nat. Hazards Earth Syst. Sci.*, vol. 16, no. 2, pp. 349–369, 2016, doi: 10.5194/nhess-16-349-2016.

[29] G. Zhao, B. Pang, Z. Xu, D. Peng, and L. Xu, “Assessment of urban flood susceptibility using semi-supervised machine learning model,” *Sci. Total Environ.*, vol. 659, pp. 940–949, 2019, doi: 10.1016/j.scitotenv.2018.12.217.

[30] M. S. Tehrany, B. Pradhan, S. Mansor, and N. Ahmad, “Flood susceptibility assessment using GIS-based support vector machine model with different kernel types,” *Catena*, vol. 125, pp. 91–101, 2015, doi: 10.1016/j.catena.2014.10.017.

[31] G. Zhao, B. Pang, Z. Xu, D. Peng, and D. Zuo, “Urban flood susceptibility assessment based on convolutional neural networks,” *J. Hydrol.*, vol. 590, no. February, p.

125235, 2020, doi: 10.1016/j.jhydrol.2020.125235.

[32] M. Rahman *et al.*, “Flood Susceptibility Assessment in Bangladesh Using Machine Learning and Multi-criteria Decision Analysis,” *Earth Syst. Environ.*, vol. 3, no. 3, pp. 585–601, 2019, doi: 10.1007/s41748-019-00123-y.

[33] E. Dodangeh *et al.*, “Integrated machine learning methods with resampling algorithms for flood susceptibility prediction,” *Sci. Total Environ.*, vol. 705, 2020, doi: 10.1016/j.scitotenv.2019.135983.

[34] K. Khosravi *et al.*, “A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran,” *Sci. Total Environ.*, vol. 627, pp. 744–755, 2018, doi: 10.1016/j.scitotenv.2018.01.266.

[35] M. S. Tehrany, S. Jones, and F. Shabani, “Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques,” *Catena*, vol. 175, no. December 2018, pp. 174–192, 2019, doi: 10.1016/j.catena.2018.12.011.

[36] S. H. Mahmoud and T. Y. Gan, “Urbanization and climate change implications in flood risk management: Developing an efficient decision support system for flood susceptibility mapping,” *Sci. Total Environ.*, vol. 636, pp. 152–167, 2018, doi: 10.1016/j.scitotenv.2018.04.282.

[37] H. Shahabi, A. Shirzadi, K. Ghaderi, and E. Omidvar, “Flood Detection and Susceptibility Mapping Using Sentinel-1 Remote Sensing Data and a Machine Learning Approach : Hybrid Intelligence of Bagging Ensemble Based on K-Nearest Neighbor Classifier,” *MDPI*, 2020, doi: Remote Sens. 2020, 12, 266; doi:10.3390/rs12020266.

[38] G. Zhao, B. Pang, Z. Xu, J. Yue, and T. Tu, “Mapping flood susceptibility in mountainous areas on a national scale in China,” *Sci. Total Environ.*, vol. 615, pp. 1133–1142, 2018, doi: 10.1016/j.scitotenv.2017.10.037.

[39] H. Darabi, B. Choubin, O. Rahmati, A. Torabi Haghighi, B. Pradhan, and B. Kløve, “Urban flood risk mapping using the GARP and QUEST models: A comparative study of machine learning techniques,” *J. Hydrol.*, vol. 569, no. December 2018, pp. 142–154, 2019, doi: 10.1016/j.jhydrol.2018.12.002.

[40] A. Arabameri *et al.*, “Modeling spatial flood using novel ensemble artificial intelligence approaches in northern Iran,” *Remote Sens.*, vol. 12, no. 20, pp. 1–30, 2020, doi: 10.3390/rs12203423.

[41] C. Cao, P. Xu, Y. Wang, J. Chen, L. Zheng, and C. Niu, “Flash flood hazard susceptibility mapping using frequency ratio and statistical index methods in

coalmine subsidence areas," *Sustain.*, vol. 8, no. 9, 2016, doi: 10.3390/su8090948.

[42] K. Uddin, M. A. Matin, and F. J. Meyer, "Operational flood mapping using multi-temporal Sentinel-1 SAR images: A case study from Bangladesh," *Remote Sens.*, vol. 11, no. 13, 2019, doi: 10.3390/rs11131581.

[43] F. Fotovatikhah, M. Herrera, S. Shamshirband, K. W. Chau, S. F. Ardabili, and M. J. Piran, "Survey of computational intelligence as basis to big flood management: Challenges, research directions and future work," *Eng. Appl. Comput. Fluid Mech.*, vol. 12, no. 1, pp. 411–437, 2018, doi: 10.1080/19942060.2018.1448896.

[44] M. Vojtek and J. Vojteková, "Flood susceptibility mapping on a national scale in Slovakia using the analytical hierarchy process," *Water (Switzerland)*, vol. 11, no. 2, 2019, doi: 10.3390/w11020364.

[45] B. Choubin, E. Moradi, M. Golshan, J. Adamowski, F. Sajedi-Hosseini, and A. Mosavi, "An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines," *Sci. Total Environ.*, vol. 651, pp. 2087–2096, 2019, doi: 10.1016/j.scitotenv.2018.10.064.

[46] A. J. Echendu, "The impact of flooding on Nigeria's sustainable development goals (SDGs)," *Ecosyst. Heal. Sustain.*, vol. 6, no. 1, 2020, doi: 10.1080/20964129.2020.1791735.

[47] O. Rahmati, H. Zeinivand, and M. Besharat, "Flood hazard zoning in Yasooj region, Iran, using GIS and multi-criteria decision analysis," *Geomatics, Nat. Hazards Risk*, vol. 7, no. 3, pp. 1000–1017, 2016, doi: 10.1080/19475705.2015.1045043.

[48] S. F. Balica, I. Popescu, L. Beevers, and N. G. Wright, "Parametric and physically based modelling techniques for flood risk and vulnerability assessment: A comparison," *Environ. Model. Softw.*, vol. 41, pp. 84–92, 2013, doi: 10.1016/j.envsoft.2012.11.002.

[49] O. Rahmati, H. R. Pourghasemi, and H. Zeinivand, "Flood susceptibility mapping using frequency ratio and weights-of-evidence models in the Golastan Province, Iran," *Geocarto Int.*, vol. 31, no. 1, pp. 42–70, 2016, doi: 10.1080/10106049.2015.1041559.

[50] M. Shafapour Tehrany, F. Shabani, M. Neamah Jebur, H. Hong, W. Chen, and X. Xie, "GIS-based spatial prediction of flood prone areas using standalone frequency ratio, logistic regression, weight of evidence and their ensemble techniques," *Geomatics, Nat. Hazards Risk*, vol. 8, no. 2, pp. 1538–1561, 2017, doi: 10.1080/19475705.2017.1362038.

[51] R. Costache and D. Tien Bui, “Spatial prediction of flood potential using new ensembles of bivariate statistics and artificial intelligence: A case study at the Putna river catchment of Romania,” *Sci. Total Environ.*, vol. 691, pp. 1098–1118, 2019, doi: 10.1016/j.scitotenv.2019.07.197.

[52] R. Costache and D. Tien Bui, “Identification of areas prone to flash-flood phenomena using multiple-criteria decision-making, bivariate statistics, machine learning and their ensembles,” *Sci. Total Environ.*, vol. 712, 2020, doi: 10.1016/j.scitotenv.2019.136492.

[53] H. Hong, W. Chen, C. Xu, A. M. Youssef, B. Pradhan, and D. Tien Bui, “Rainfall-induced landslide susceptibility assessment at the Chongren area (China) using frequency ratio, certainty factor, and index of entropy,” *Geocarto Int.*, vol. 32, no. 2, pp. 139–154, 2017, doi: 10.1080/10106049.2015.1130086.

[54] Y. Wang *et al.*, “Flood susceptibility mapping in Dingnan County (China) using adaptive neuro-fuzzy inference system with biogeography based optimization and imperialistic competitive algorithm,” *J. Environ. Manage.*, vol. 247, no. July, pp. 712–729, 2019, doi: 10.1016/j.jenvman.2019.06.102.

[55] A. Mosavi, P. Ozturk, and K. W. Chau, “Flood prediction using machine learning models: Literature review,” *Water (Switzerland)*, vol. 10, no. 11, pp. 1–40, 2018, doi: 10.3390/w10111536.

[56] M. B. Kia, S. Pirasteh, B. Pradhan, A. R. Mahmud, W. N. A. Sulaiman, and A. Moradi, “An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia,” *Environ. Earth Sci.*, vol. 67, no. 1, pp. 251–264, 2012, doi: 10.1007/s12665-011-1504-z.

[57] S. Lee and I. Park, “Application of decision tree model for the ground subsidence hazard mapping near abandoned underground coal mines,” *J. Environ. Manage.*, vol. 127, pp. 166–176, 2013, doi: 10.1016/j.jenvman.2013.04.010.

[58] S. Lee *et al.*, “Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city , Korea,” *Sci. Total Environ.*, vol. 5705, p. 20, 2017, doi: 10.1080/19475705.2017.1308971.

[59] Y. Wang *et al.*, “A hybrid GIS multi-criteria decision-making method for flood susceptibility mapping at Shangyou, China,” *Remote Sens.*, vol. 11, no. 1, 2019, doi: 10.3390/rs11010062.

[60] B. Ayo, “INTEGRATING OPENSTREETMAP DATA AND SENTINEL-2 IMAGERY FOR CLASSIFYING AND MONITORING INFORMAL SETTLEMENTS,” 2020.

[61] P. Data, “City Population,” 2020. http://www.citypopulation.de/ (accessed Sep. 14,

2020).

[62] U. of E. Anglia, “Rainfall Data,” 2020. http://www.cru.uea.ac.uk/data (accessed Sep. 18, 2020).

[63] O. S. Map, “Road Data obtained from Geofabrik Website.” http://www.geofabrik.de/ (accessed Oct. 02, 2020).

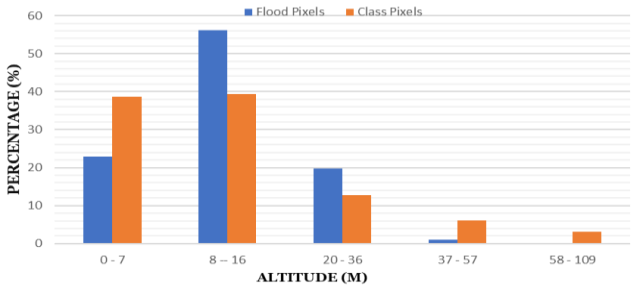
[64] U. C. Nkwunonwo, F. I. Okeke, E. S. Ebinne, and N. E. Chiemelu, “Free, open, quantitative and adaptable digital soil map data and database for Nigeria,” *Data Br.*, vol. 31, p. 105941, 2020, doi: 10.1016/j.dib.2020.105941.

[65] M. S. Tehrany, B. Pradhan, and M. N. Jebur, “Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS,” *J. Hydrol.*, vol. 512, no. May, pp. 332–343, 2014, doi: 10.1016/j.jhydrol.2014.03.008.

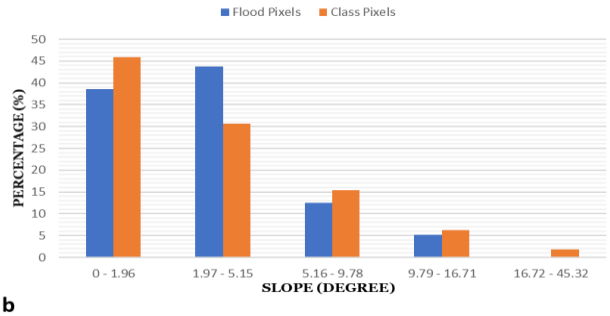
[66] H. Mojaddadi, B. Pradhan, H. Nampak, N. Ahmad, and A. H. bin Ghazali, “Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and GIS,” *Geomatics, Nat. Hazards Risk*, vol. 8, no. 2, pp. 1080–1102, 2017, doi: 10.1080/19475705.2017.1294113.

9 ANNEXES

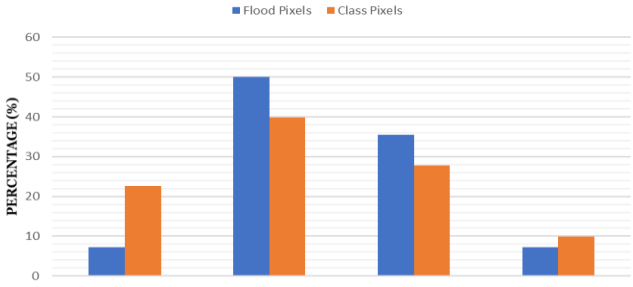
9.1 Relative Distribution of flood pixels within Flood Influencing Factors' Classes

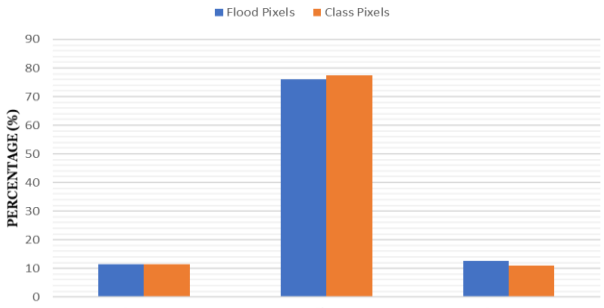


a

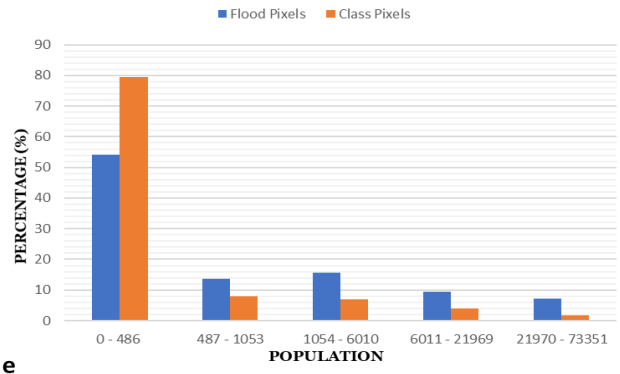


b

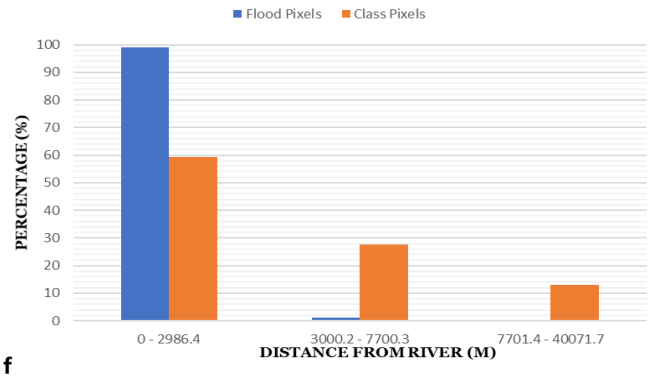




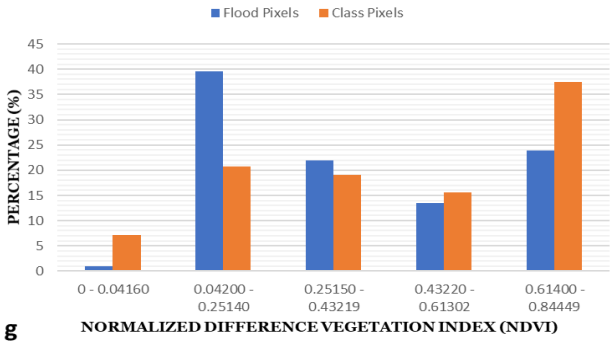
d



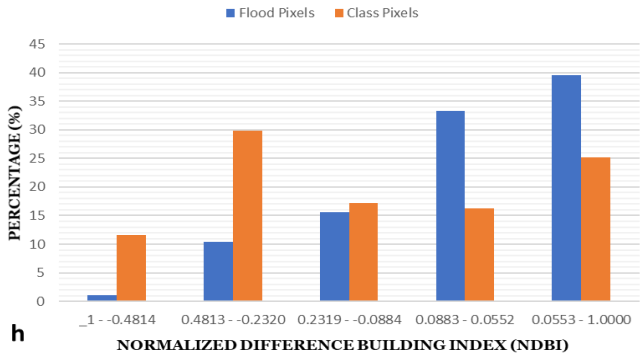
e



f

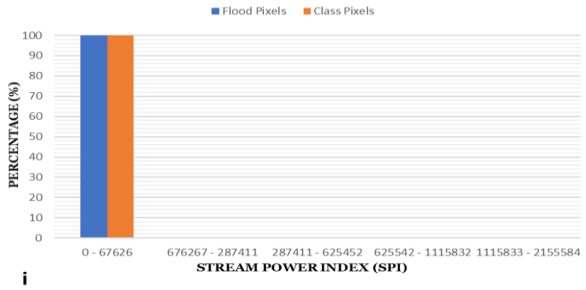


g

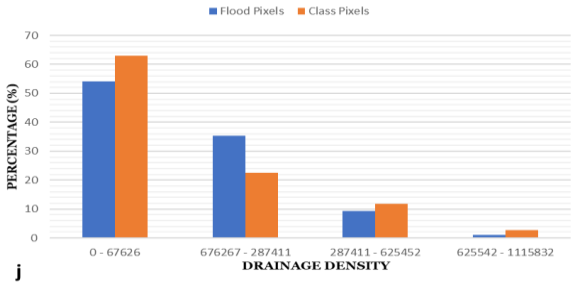


h

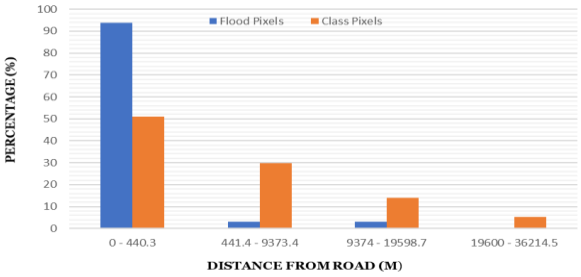
Figure 9. 1: (a) Altitude, (b)Slope, (c) Rainfall, (d) Curvature, (e) Population density, (f) Distance from River, (g) NDVI, (h) NDBI, (i) SPI, (j) Drainage Density, (k) Distance from Road, (l) TWI, (m) LULC, (n) Geology, (o) Soil



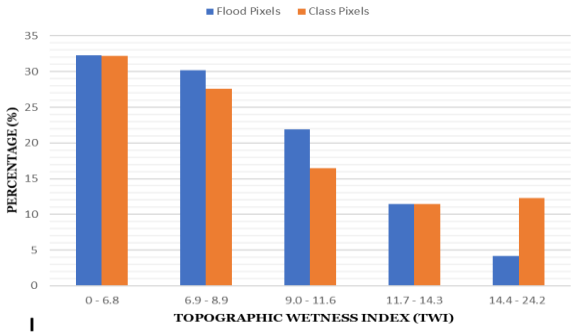
i



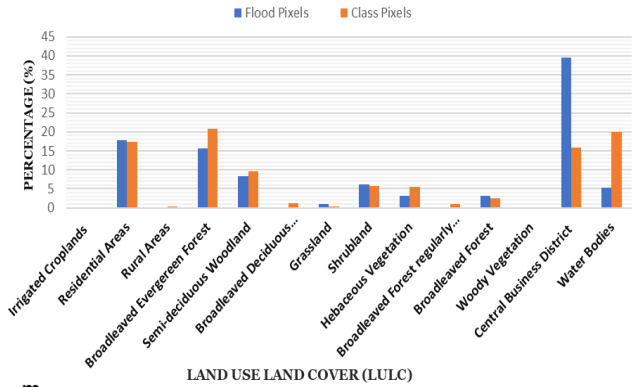
j



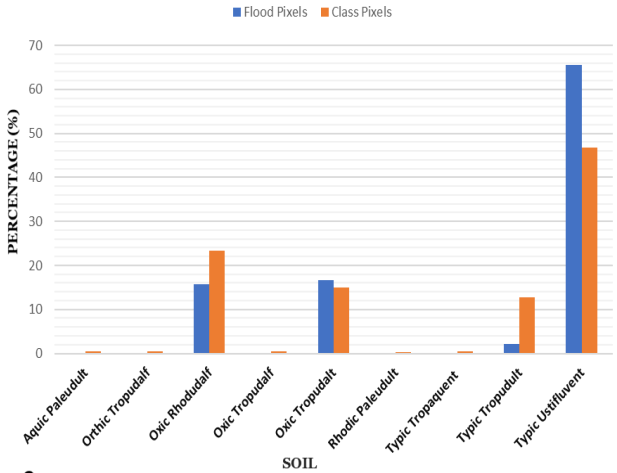
k



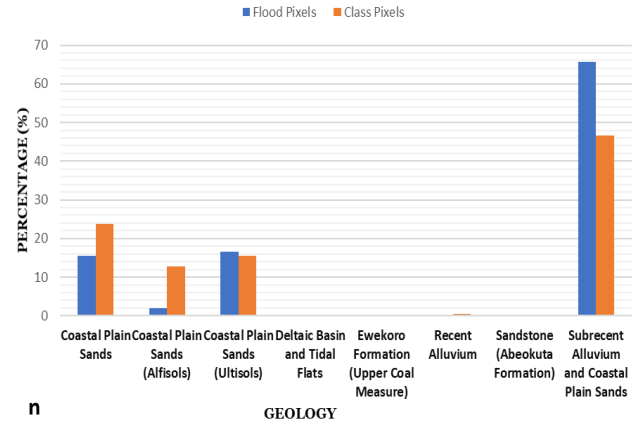
l



m



o



n

9.2 Descriptive statistics of the training and testing datasets

		Rainfall	Altitude	Curvature	Slope	TWI	SPI	Drainage Density	LULC	Soil	Geology	Population	Dist_F_River	Dist_F_Road	NDVI	NDBI
Training Sample	**Mean**	10.53	16.35	-1043.45	3.99	7.20	7.87	4234.98	107.70	7.01	5.19	8160.86	4.86	45.42	0.43	-0.06
	S.E	0.07	1.12	774.16	0.33	0.09	2.37	408.87	5.24	0.18	0.22	947.34	0.43	4.52	0.02	0.01
	Median	10.44	11.00	0.00	2.32	7.18	0.00	1676.86	120.00	9.00	8.00	1352.07	2.67	25.48	0.39	-0.02
	S.D	0.98	15.59	10810.57	4.54	1.32	33.06	5709.54	73.14	2.52	3.11	13228.92	6.06	63.11	0.24	0.19
	S.V	0.97	243.02	116868329.51	20.63	1.74	1093.00	32598897.86	5349.09	6.37	9.70	175004261.89	36.68	3982.77	0.06	0.04
	Kurtosis	0.44	4.37	3.50	4.96	0.86	121.58	6.05	-1.82	-1.30	-1.81	2.06	1.36	6.97	-1.39	0.19
	Skewness	0.50	2.04	-0.12	2.12	0.64	10.25	1.99	0.15	-0.70	-0.28	1.62	1.45	2.40	-0.03	-0.69
	Range	5.03	87.00	84240.00	23.52	7.91	415.48	37342.90	180.00	7.00	7.00	69157.13	24.88	350.84	0.94	1.00
	Minimum	8.36	0.00	-44064.00	0.00	4.30	0.00	0.00	30.00	2.00	1.00	-4843.43	0.00	0.00	-0.16	-0.73
	Maximum	13.38	87.00	40176.00	23.52	12.21	415.48	37342.90	210.00	9.00	8.00	64313.70	24.88	350.84	0.78	0.27
	Sum	2052.49	3189.00	-203472.00	777.52	1403.99	1535.34	825821.16	21002.00	1367.00	1012.00	1591367.39	948.48	8856.17	82.91	-12.23
	Count	195	195	195	195	195	195	195	195	195	195	195	195	195	195	195
Testing Sample	**Mean**	10.35	14.66	93.69	3.26	7.42	8.45	3145.15	116.86	6.90	5.06	7663.99	3.51	30.03	0.41	-0.07
	S.E	0.09	1.42	961.43	0.35	0.17	2.90	487.07	8.20	0.29	0.36	1326.29	0.55	4.96	0.02	0.02
	Median	10.26	11.00	0.00	2.10	7.12	0.00	426.27	120.00	9.00	8.00	2207.44	1.91	18.72	0.36	-0.01
	S.D	0.81	12.95	8759.05	3.19	1.57	26.39	4437.44	74.71	2.67	3.26	12083.12	5.03	45.16	0.23	0.20
	S.V	0.66	167.69	76720890.58	10.16	2.47	696.39	19690874.59	5581.08	7.11	10.64	146001827.02	25.31	2039.74	0.05	0.04
	Kurtosis	0.41	4.95	5.02	4.95	0.52	25.96	1.62	-1.87	-1.48	-1.90	2.54	3.69	18.43	-1.50	0.92
	Skewness	0.70	2.11	0.47	2.13	1.02	4.82	1.49	-0.09	-0.65	-0.25	1.72	2.00	3.50	0.12	-1.03
	Range	3.90	65.00	67392.00	14.71	6.75	181.69	18192.50	180.00	6.00	7.00	54556.57	21.79	313.87	0.70	0.94
	Minimum	9.06	0.00	-28512.00	0.00	5.34	0.00	0.00	30.00	3.00	1.00	-4455.17	0.00	0.00	0.07	-0.70
	Maximum	12.95	65.00	38880.00	14.71	12.09	181.69	18192.50	210.00	9.00	8.00	50101.40	21.79	313.87	0.77	0.24
	Sum	859.06	1217.00	7776.00	270.33	616.07	701.69	261047.67	9699.00	573.00	420.00	636111.00	291.19	2492.82	33.87	-5.87
	Count	83	83	83	83	83	83	83	83	83	83	83	83	83	83	83

S.E – Standard Error; S.D – Standard Deviation; S.V – Sample Variance.

Masters Program in Geospatial Technologies

Supported by:

Education and Culture

ERASMUS MUNDUS